Italian Statistical Society Series on Advances in Statistics

Alessio Pollice · Paolo Mariani *Editors*

Methodological and Applied Statistics and Demography I

SIS 2024, Short Papers, Plenary and Specialized Sessions

SIS Società Italiana di Statistica

Springer

Italian Statistical Society Series on Advances in Statistics

This book series publishes volumes of peer-reviewed short papers presented at the scientific events such as conferences, seminars and workshops organized by the Italian Statistical Society (SIS) and its sections.

The Italian Statistical Society (Società Italiana di Statistica, SIS) was established in 1939 and ranks among the institutions of particular scientific relevance. SIS aims to promote scientific activities for the development of statistical sciences and carries out this task organizing scientific meetings and conferences, and by means of publications and partnerships at a national and international level.

Alessio Pollice · Paolo Mariani
Editors

Methodological and Applied Statistics and Demography I

SIS 2024, Short Papers, Plenary and Specialized Sessions

Springer

Editors
Alessio Pollice
Department of Economics and Finance
The University of Bari Aldo Moro
Bari, Italy

Paolo Mariani
Department of Economics, Management
and Statistics
University of Milano-Bicocca
Milan, Italy

ISSN 3059-2135 ISSN 3059-2143 (electronic)
Italian Statistical Society Series on Advances in Statistics
ISBN 978-3-031-64345-3 ISBN 978-3-031-64346-0 (eBook)
https://doi.org/10.1007/978-3-031-64346-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The promotion of scientific activities for the advancement of statistical sciences represents the defining element of the Italian Statistical Society's (SIS) actions. Since its inception, the Society has grown in tandem with the rapid dissemination of quantitative methods for data analysis across various fields of scientific research and social life. Academics, researchers, doctoral students, and professionals interested in statistical and demographic research, as well as related methodological and applied fields, have shared their activities during the 52nd Scientific Meeting of the Italian Statistical Society held at the University of Bari Aldo Moro from June 17th to June 20th, 2024. The biennial conference is a traditional event that promotes interactions among national and international researchers in Statistics, Demography, and applied Statistics in Italy. The objective of the conference is to bring together national and foreign researchers and professionals to discuss recent developments in theoretical and applied Statistics, as well as Demography and Statistics for the social sciences. The event featured over 450 presentations, including 7 keynotes in the 5 plenary sessions, 66 contributions in the 22 specialized sessions, 144 contributions in the 36 invited sessions, and 236 presentations in the 48 sessions of spontaneous communications, all addressing specific themes of methodological and/or applied Statistics and Demography. The high number of contributions and the significant participation in the conference demonstrate the desire to participate in scientific events as venues for exchange and discussion on the new developments in our field, a vibrant characteristic of our scientific community.

Round tables and meetings on topics of general interest such as "Data science: statistics for the 22nd century" and "Policy tools for innovation and development" were also organized, along with satellite events such as the 7th edition of the hackathon "Stats Under the Stars".

We would also like to thank the European Network for Business and Industrial Statistics (ENBIS) and the Institute of Mathematical Statistics (IMS) for their presence. Special thanks go to the Program and Organizing Committees for their tremendous effort, to the University of Bari Aldo Moro for endorsing and supporting the event, and to all the sponsors for their contributions: Società Italiana di Statistica, Università degli Studi di Bari Aldo Moro, Banca Popolare di Puglia e Basilicata, Fondazione Puglia, Ladisa Ristorazione S.R.L., and Unioncamere Puglia. The high number of contributions has led to the production of four volumes based on the types of sessions outlined in the scientific program. The first volume includes the Plenary and Specialized sessions, the second volume includes the Solicited sessions, and the third and fourth volumes include the Contributed sessions. These volumes contain works, as usual subjected to peer review, that the authors have agreed to publish, providing a comprehensive overview of statistical research in recent years.

This first volume presents a selection of contributions from authors who delivered invited presentations during the Plenary and Specialized sessions.

June 2024

Alessio Pollice
Paolo Mariani

Organization

SIS 2024 is organized by the Department of Economics and Finance, University of Bari Aldo Moro (UniBA) in cooperation with the Italian Statistical Society.

Editors

Alessio Pollice	University of Bari Aldo Moro
Paolo Mariani	University of Milano, Bicocca

Executive Editors

Diego Battagliese	University of Bari Aldo Moro
Massimo Bilancia	University of Bari Aldo Moro
Crescenza Calculli	University of Bari Aldo Moro
Vittorio Nicolardi	University of Bari Aldo Moro

Program Committee Chair

Paolo Mariani	University of Milano, Bicocca

Program Committee Members

Elena Ambrosetti	University of Roma, Sapienza
Laura Anderlucci	University of Bologna
Margaret Antonicelli	University of Roma, Sapienza
Raffaele Argiento	University of Bergamo
Maria Gabriella Campolo	University of Messina
Maurizio Carpita	University of Brescia
Enrico Ciavolino	University of Salento
Andrea Ciccarelli	University of Teramo
Enrico Di Bella	University of Genova
Giuseppe Gabrielli	University of Napoli, Federico II
Salvatore Ingrassia	University of Catania
Antonio Irpino	University of Campania
Fulvia Mecatti	University of Milano, Bicocca

Stefano Menghiniello	Italian National Institute of Statistics
Alessio Pollice	University of Bari Aldo Moro
Mariangela Sciandra	University of Palermo
Giorgia Zaccaria	University of Roma, Sapienza

Organizing Committee Chair

Alessio Pollice	University of Bari Aldo Moro

Organizing Committee Members

Diego Battagliese	University of Bari Aldo Moro
Antonella Bianchino	Italian National Institute of Statistics
Massimo Bilancia	University of Bari Aldo Moro
Crescenza Calculli	University of Bari Aldo Moro
Francesco Campobasso	University of Bari Aldo Moro
Maria Carella	University of Bari Aldo Moro
Carlo Cusatelli	University of Bari Aldo Moro
Francesco Domenico D'Ovidio	University of Bari Aldo Moro
Angela Maria D'Uggento	University of Bari Aldo Moro
Najada Firza	University of Bari Aldo Moro
Michele Gallo	University of Napoli, L'Orientale
Thais Garcia Pereiro	University of Bari Aldo Moro
Samuela L'Abbate	University of Bari Aldo Moro
Fabio Manca	University of Bari Aldo Moro
Claudia Marin	University of Bari Aldo Moro
Caterina Marini	University of Bari Aldo Moro
Antonella Massari	University of Bari Aldo Moro
Rocco Mazza	University of Bari Aldo Moro
Dante Mazzitelli	University of Bari Aldo Moro
Roberta Misuraca	University of Bari Aldo Moro
Vittorio Nicolardi	University of Bari Aldo Moro
Roberta Pace	University of Bari Aldo Moro
Anna Paterno	University of Bari Aldo Moro
Michela Camilla Pellicani	University of Bari Aldo Moro
Paola Perchinunno	University of Bari Aldo Moro
Nunziata Ribecco	University of Bari Aldo Moro
Ernesto Toma	University of Bari Aldo Moro
Giovanni Vannella	University of Bari Aldo Moro
Domenico Viola	University of Bari Aldo Moro

Referees

G. Adelfio
L. S. Alaimo
A. Albano
G. Alderotti
M. Angelelli
R. Argiento
S. Arima
F. Ascolani
D. Battagliese
F. Benassi
M. Beraha
R. Berni
C. Bernini
G. Bertarelli
I. Bombelli
C. Bonifazi
A. Busetta
G. Busetta
A. Calcagni
C. Calculli
M. Cameletti
F. Camerlenghi
M. G. Campolo
A. Canale
C. Capezza
A. Cappozzo
M. Carpita
A. Casa
A. Cassese
C. Castagnaro
C. Cavicchia
R. Cerqueti
A. Ciccarelli
C. Conversano
A. D'Ambra
A. D'Ambrosio
C. Di Serio
S. Di Zio
F. D'Ovidio
A. M. D'Uggento
E. Fabrizi
F. Finazzi
S. Franceschi
G. Gabrielli
M. Gallo
G. Ghellini
A. Ghiglietti
F. A. Giambona
C. Gigliarano
S. Giordano
C. Gozzi
A. Guglielmi
M. Iannario
R. Ignaccolo
R. Leombruni
C. Liberati
B. Liseo
R. Lombardo
A. Lucadamo
F. Maggino
C. Marin
A. Marletta
F. Martella
G. Mastrantonio
A. Mattei
A. Mazza
S. Menghinello
C. Mollica
A. Naccarato
M. Nai Ruscone
B. Nipoti
M. Novelli
R. Pace
S. Pacei
L. Paci
A. Paparusso
F. Pavone
C. Polli
A. Pollice
M. Pratesi
I. Prosdocimi
G. Punzo
G. Ragozini
F. Rapallo
A. Rinaldi
E. Ripamonti

E. Romano
S. Salini
L. Sangalli
P. Sarnacchiaro
M. Sciandra
A. Simonetto
S. D. Tomarchio
V. Tomaselli
E. Trappolini
C. Valentino
R. Verde
P. Vicard
C. Viroli
M. P. Vitale
M. Zenga

Sponsoring Institutions

Società Italiana di Statistica, Roma, Italy
Università degli Studi di Bari Aldo Moro, Bari, Italy
Banca Popolare di Puglia e Basilicata, Altamura (BA), Italy
Fondazione Puglia, Bari, Italy
Ladisa Ristorazione S.R.L., Bari, Italy
Unioncamere Puglia, Bari, Italy

List of Papers

Plenary Sessions

- Daniele Durante, Conjugacy and approximation properties of skew-symmetric distributions in statistical inference
- Francesca Ieva, Statistical Learning in healthcare: towards a new paradygm of research
- Janine Illian, Realistically complex spatio-temporal modelling - responsibilities, challenges and fallacies

Specialized Sessions

– Advanced statistical methods for environmental sustainability

- Donato Riccio, Fabrizio Maturo, Elvira Romano, Supervised Classification of Functional Data via Ensembles of Different Functional Representations
- Elvira Romano, Andrea Diana, Anna De Magistris, Conformal based uncertainty bands for predictions in functional ordinary kriging

– Advanced Statistical Models for Data Science

- Alessandro Albano, Chiara Di Maria, Mariangela Sciandra, Antonella Plaia, Causal machine learning for medical texts

– Advances on the analysis of preference data

- José Luis García-Lapresta, Miguel Martínez-Panero, A positional analysis of ranking procedures
- Marta Nai Ruscone, Antonio D'Ambrosio, Non-metric unfolding via copula
- Rosaria Simone, Analyzing differences in ranking distributions

– Bayesian statistics for real-world application: challenges and solution

- Aldo Gardini, Silvia De Nicolò, Enrico Fabrizi, Mapping well-being through a Mixture-of-Experts Fay-Herriot model

– Data eco-systems and the value of information

- Gianna Greca, Alessandro Zeli, The design of an energy data eco-system. An analysis of user-specific information needs and data integration processes.
- Francesco Pugliese, Angela Pappagallo, Massimo De Cubellis, Generation of Synthetic Data from Mobile Network Operators (MNO) data through Generative Adversarial Networks (GANs)

- Nonparametric causal inference
  - Dafne Zorzetto, Causal STAR BART for discrete outcome
- Re-positioning sampling statistics in the multi data-source world
  - Elisabetta Carfagna, Data science, citizen science and smart official statistics
  - Cristiano Ferraz, Sampling challenges from the 2030 agricultural sustainable development goals
  - Pier Luigi Conti, Methodologcal perspectives in integration of data from multiple probabilistic and non-probabilistic sources
- Residential patterns of foreign migrants in urban contexts
  - Francesca Bitonti, Daniela Ghio, Angelo Mazza, Massimo Mucciardi, An Information Theory approach to assess residential segregation: the case of Messina (Italy)
  - Alessio Buonomo, Federico Benassi, Rosaria Simone, Salvatore Strozza, Multiscale dimensions of residential segregation in Naples. A preliminary investigation
  - Maria Miriam Carella, Thais Garcia Pereiro, Anna Paterno, Socioeconomic distress and foreign presence in a Southern urban context. The case of Bari
- Spatio-temporal methods for environmental sustainability
  - Julia Jansson, Christophe A. N. Biscio, Mehdi Moradi, Ottmar Cronie, Point Process Learning: a cross-validation-based statistical framework for point processes
  - Marco Mingione, Pierfrancesco Alaimo Di Loro, Temporal Nearest Neighbor Gaussian Process (tNNGP) with flexible covariance for modelling physical activity
  - Florian Wolf, Alessandro Carminati, Alessandra Guglielmi, Spatio-temporal clustering of PM2.5 in northern Italy using a Bayesian model
- Statistical Analysis of Economic Complexity
  - Cristina Martelli, Adham Kahlawi, Maria Flora Salvatori, Graphs as Unifying Logical Structures in the Construction of Information System for complex domains
  - Furio Urso, Antonino Abbruzzo, Marcello Chiodi, Maria Francesca Cracolici, Clickstream Data Analysis and Web User Profiling via Mixture Hidden Markov Models
- Statistical approaches for analysing social data on violence against women
  - Fiorenza Deriu, Emilia La Nave, Analysing language for preventing women from gender violence: NLP and Machine Learning techniques to classify Tweet messages
  - Silvia Polettini, Sara Martino, Greta Panunzi, A Pogit model for underreported counts of violence against women in Italy
- Statistical modeling for environmental quality and biodiversity
  - Antonella Congedi, Sandra De Iaco, A dynamic spatial indicator of the surface water quality

Contents

Plenary Sessions

Conjugacy and Approximation Properties of Skew–Symmetric Distributions in Statistical Inference

Daniele Durante(✉)

Department of Decision Sciences, Bocconi University, Via Röntgen 1, 20136 Milan, Italy
daniele.durante@unibocconi.it

Abstract. Among the several areas of Statistics that have witnessed foundational contributions by the Italian community, a remarkable one relates to the development and study of tractable classes of skewed distributions. These families arise from the very elegant and practical idea of perturbing symmetric densities through a skewness–inducing mechanism that preserves tractability of the resulting class of distributions. As such, these families have been widely and successfully adopted in the design of statistical models for skewed phenomena. This short article aims at clarifying that the impact of skewed distributions goes even beyond the specification of effective likelihoods. In fact, within these families it is possible to identify (i) the conjugate priors for a broad variety of core statistical models often employed in practice (i.e., linear regression, probit, tobit, and multinomial probit), and (ii) novel limiting laws for generic posterior distributions that substantially improve the rate of convergence, and hence the approximation accuracy, relative to those provided by the classical Gaussians from the Laplace method. Such results are presented in this short article through a summary of the contributions by [1,15,16,18,20].

Keywords: Bayesian inference · Skew–symmetric distribution · Unified skew–elliptical distribution · Unified skew–normal distribution

1 Skewed Perturbations of Symmetric Densities

A celebrated artist, Édouard Manet, once claimed: *"There's no symmetry in nature. One eye is never exactly the same as the other. There's always a difference"*. When I first came across skew–normal distributions [8,11] towards the end of my studies as a student in Statistics at the University of Padova, I realized that there was no symmetry, by definition, even in the distribution of the grade I got at the very first course in Statistics I attended, namely, Descriptive Statistics. Professor Bruno Scarpa, who became later my Ph.D. supervisor, designed the final exam for such a course in a quite peculiar manner. Students who received a grade above a pre–specified high threshold $\bar{z}$ in the written part

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 3–9, 2025.
https://doi.org/10.1007/978-3-031-64346-0_1

were required to take also an oral exam. For these students, the final grade would have been that of the oral exam, with no guarantees that such a grade were higher than the one obtained in the written part. The motivation Professor Scarpa gave us for such an unusual exam was that he wanted to meet and further test the top students in his class. In fact, I remember he was quite surprised when he saw me at the oral exam, provided that during the lectures I was not among those students in the front rows interacting constantly with him and replying to every single difficult question he asked. Quite the opposite.

Besides giving me chance to interact with my future mentor, the above exam was also the first time I unconsciously came across something similar to a skew–normal. To see why, let z_0 and z_1 be my grades at generic written and oral exams, respectively, on the topics of the course in Descriptive Statistics. Moreover, with an excess of simplicity, assume

$$\begin{bmatrix} z_1 \\ z_0 \end{bmatrix} \sim \mathrm{N}_2 \left(\begin{bmatrix} \mu_1 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{bmatrix} \right),$$

and that the threshold $\bar{z}$ set by the professor to access the oral exam was equal to μ_0. Finally, recall also that, by the closure under conditioning of Gaussian distributions, it holds $(z_0 \mid z_1) \sim \mathrm{N}(\mu_0 + [\rho\sigma_0/\sigma_1](z_1 - \mu_1), (1-\rho^2)\sigma_0^2)$. Then by the structure of the exam, once admitted to the oral part, my final grade would have been distributed as $(z_1 \mid z_0 > \bar{z})$ with density

$$\begin{aligned} p(z_1 \mid z_0 > \bar{z}) &= p(z_1 \mid z_0 > \mu_0) = p(z_1) \frac{\mathrm{pr}(z_0 > \mu_0 \mid z_1)}{\mathrm{pr}(z_0 > \mu_0)} \\ &= 2\phi(z_1 - \mu_1; \sigma_1^2) \Phi([\rho\sigma_0/\sigma_1](z_1 - \mu_1); (1-\rho^2)\sigma_0^2), \end{aligned}$$

where the generic quantities $\phi(z; \sigma^2)$ and $\Phi(z; \sigma^2)$ denote, respectively, the density and cumulative distribution function, evaluated at z, of the univariate Gaussian with mean 0 and variance σ^2. Comparing the above expression with those in Chap. 2 of the book "The Skew–Normal and Related Families" by Azzalini and Capitanio [11], it is clear that, up to a reparameterization, $(z_1 \mid z_0 > \bar{z})$ is distributed as a univariate skew–normal. Such a connection clarifies that, although z_1 is marginally symmetric, once I had been admitted to the oral part according to the exam rules, the distribution of my final grade suddenly inherited a skewed behavior regulated by the perturbation factor $\Phi([\rho\sigma_0/\sigma_1](z_1 - \mu_1); (1-\rho^2)\sigma_0^2)$. Its effect, was that of inflating the unconditional Gaussian density $p(z_1)$ above μ_1 while deflating the part below μ_1, when $\rho > 0$. The opposite happens if $\rho < 0$.

The above example clarifies that skewed perturbations of symmetric densities are ubiquitous in practice, while being inherent to phenomena arising from selection representations [3]. Such a practical relevance has led to important generalizations of the original univariate skew–normal distribution [8] to account for (i) perturbation of more general symmetric densities beyond the Gaussian one [3,4,10], (ii) additional flexibility in the truncation mechanisms [6,7], and (iii) multivariate representations, both for the original density to be perturbed and for the skewing factor [2,9,12,19,21]. Among all these extensions,

two remarkable ones for generality and properties are the unified skew–normal (SUN) distribution [2], and the flexible skew–symmetric (SKS) class [10,24].

The SUN family has been originally introduced in [2] in order to unify within a single representation several extensions of the original multivariate skew–normal developed by [12]. Under such a unification, a generic random vector $\mathbf{z} \in \mathbb{R}^p$ has $\text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\gamma}, \boldsymbol{\Delta}, \boldsymbol{\Gamma})$ distribution if its density can be expressed as

$$p(\mathbf{z}) = \phi_p(\mathbf{z}-\boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_q[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\mathbf{z}-\boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]}{\Phi_q(\boldsymbol{\gamma}; \boldsymbol{\Gamma})}, \tag{1}$$

where $\phi_p(\mathbf{z}-\boldsymbol{\xi}; \boldsymbol{\Omega})$ denotes the density, evaluated at $\mathbf{z}-\boldsymbol{\xi}$, of a $\text{N}_p(\mathbf{0}, \boldsymbol{\Omega})$, whereas $\Phi_q[\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\mathbf{z}-\boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta}]$ and $\Phi_q(\boldsymbol{\gamma}; \boldsymbol{\Gamma})$ correspond to the cumulative distribution functions of the q–variate Gaussians $\text{N}_q(\mathbf{0}, \boldsymbol{\Gamma} - \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})$ and $\text{N}_q(\mathbf{0}, \boldsymbol{\Gamma})$, computed at $\boldsymbol{\gamma} + \boldsymbol{\Delta}^\top \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\mathbf{z}-\boldsymbol{\xi})$ and $\boldsymbol{\gamma}$, respectively, with $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1} \boldsymbol{\Omega} \boldsymbol{\omega}^{-1}$. As is clear from (1), such a class extends the univariate skew–normal in order to account for perturbation of generic p–dimensional Gaussian densities with respect to skewness inducing mechanisms regulated by cumulative distribution functions of multivariate normals. The magnitude of such a perturbation is mainly regulated by $\boldsymbol{\Delta} \in \mathbb{R}^{p \times q}$. When $\boldsymbol{\Delta} = \mathbf{0}$, (1) reduces to $\phi_p(\mathbf{z}-\boldsymbol{\xi}; \boldsymbol{\Omega})$.

Besides yielding a wide family of distributions for modeling skewed phenomena, the above class crucially preserves the closure under linear combination, conditioning and marginalization of the original Gaussian family. In addition, it admits a closed–form expression for the moment generating function along with a simple stochastic representation via linear combinations of p–dimensional Gaussians and q–variate truncated normals. This tractability has motivated extensive studies of SUNs and rapid extensions to other flexible classes. A noticeable advancement in this direction is provided by the skew–symmetric (SKS) family [10,24]. In particular, to continue the progression towards increasingly–flexible formulations, it is useful to focus, in our case, on the representation provided by [24] which is guaranteed to recover some instances of SUNs as a special case. Under this construction, a random vector $\mathbf{z} \in \mathbb{R}^p$ has SKS distribution if its density is

$$p(\mathbf{z}) = 2 f_p(\mathbf{z}-\boldsymbol{\xi}) w(\mathbf{z}-\boldsymbol{\xi}), \tag{2}$$

where $f_p(\mathbf{z}-\boldsymbol{\xi})$ denotes the density, evaluated at $\mathbf{z}-\boldsymbol{\xi}$, of a generic p–dimensional distribution symmetric at $\mathbf{0}$ (possibly indexed by additional parameters), while $w(\cdot) : \mathbb{R}^p \to [0,1]$ is a skewing function that satisfies $w(-(\mathbf{z}-\boldsymbol{\xi})) = 1 - w(\mathbf{z}-\boldsymbol{\xi})$. As proved in [24], expression (2) defines a proper density which also admits a tractable stochastic representation for $\mathbf{z}$ based on sign perturbations of samples from the symmetric component. In addition, it can be readily noticed that suitable choices of $f_p(\cdot)$ and $w(\cdot)$ allow to recover some key instances of SUNs in (1) as a special case. It shall be also emphasized that the above formulation comprises other important skewed distributions, such as some instances in the unified skew–elliptical (SUE) class [4], that can be derived by replacing $\phi_p(\cdot;\cdot)$ and $\Phi_q(\cdot;\cdot)$ in (1) with general elliptical densities and cumulative distribution functions [17].

Sections 2 and 3 show that (1)–(2) not only provide flexible families to model skewed phenomena, but also play a fundamental role in Bayesian statistics.

2 Conjugacy Properties of SUN Distributions

The intersection between skewed distributions and Bayesian statistics has witnessed interesting contributions over the past years [5,13,14,21,22]. Most of the focus has been, however, on Bayesian inference for the parameters of the skew–normal distribution and its extensions. Whether these classes can act as conjugate priors for the likelihood induced by other statistical models has remained an unaddressed question until recently. This question has, in fact, a fascinating and impactful answer. As shown in [1,15,18] SUN distributions, which include Gaussians as a special case, are the conjugate priors for the regression parameters in linear, probit, tobit and multinomial probit models, along with the corresponding extensions to multivariate, dynamic and skewed contexts.

The above result means that SUNs, and thus Gaussian, priors, lead to SUN posteriors whenever combined, via Bayes rule, with the likelihood induced by all the aforementioned models. As such, Bayesian inference under the resulting SUN posterior inherits the tractability and properties briefly discussed in Sect. 1 for this family, thereby overcoming the challenges that arise in Bayesian computation under intractable posteriors. To appreciate the beauty and simplicity of this result, consider the probit likelihood $p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \Phi(\mathbf{x}_i^\top \boldsymbol{\theta})^{y_i}[1 - \Phi(\mathbf{x}_i^\top \boldsymbol{\theta})]^{1-y_i}$, with responses $y_i \in \{0, 1\}$, covariates $\mathbf{x}_i \in \mathbb{R}^p$ and coefficients vector $\boldsymbol{\theta} \in \mathbb{R}^p$. Moreover, notice that $p(\mathbf{y} \mid \boldsymbol{\theta})$ can be equivalently rewritten as $\prod_{i=1}^{n} \Phi[(2y_i - 1)\mathbf{x}_i^\top \boldsymbol{\theta}] = \Phi_n(\mathbf{D}\mathbf{X}\boldsymbol{\theta}; \mathbf{I}_n)$, where $\mathbf{X}$ is the design matrix and $\mathbf{D} = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)$. Then, under the Gaussian prior $\boldsymbol{\theta} \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ and by the Bayes rule, we have

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}) &= \phi_p(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \Phi_n(\mathbf{D}\mathbf{X}\boldsymbol{\theta}; \mathbf{I}_n) \\ &= \phi_p(\boldsymbol{\theta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \Phi_n(\mathbf{D}\mathbf{X}\boldsymbol{\xi} + \mathbf{D}\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\xi}); \mathbf{I}_n). \end{aligned}$$

Comparing the above expression with the one in (1) (setting $\mathbf{z} = \boldsymbol{\theta}$), it is clear that $p(\boldsymbol{\theta} \mid \mathbf{y})$ is proportional to the kernel of a $\text{SUN}_{p,n}(\boldsymbol{\xi}_{\text{post}}, \boldsymbol{\Omega}_{\text{post}}, \boldsymbol{\gamma}_{\text{post}}, \boldsymbol{\Delta}_{\text{post}}, \boldsymbol{\Gamma}_{\text{post}})$ with suitably–defined parameters; see [15] for the precise expression of these parameters. As clarified in [1] a similar reasoning can be applied to prove the conjugacy of general SUN priors for the broader class of models that also include linear regression, multinomial probit and tobit, among others. See [20,23] for ongoing research on conjugacy under skew–elliptical extensions of these formulations, and binary regression models with general link functions, respectively. These clarify how such results can open several interesting directions for future research.

3 Approximation Properties of SKS Distributions

Clearly, not all posteriors are within the SUN class. Nonetheless, such a class and, more generally, the flexible SKS family, might prove helpful in deriving novel deterministic approximations of general intractable posterior distributions that improve the accuracy of the classical Gaussian one from the Laplace method. Such a latter approximation neglects terms higher than the second–order one in

the Taylor expansion of the log–posterior, provided that these higher order terms do not seem to yield kernels of known and proper densities.

In fact, [16] have recently shown that the third–order term can be accurately approximated by a suitably–defined univariate cumulative distribution function, without changing the asymptotic order of the remainder in the expansion. This results in a valid higher–order approximating density which crucially belongs to the SKS class, thereby inheriting its tractability. While [16] provide an extensive theoretical and methodological treatment of the SKS approximating class, let us focus on its practical skew–modal version. In particular, let $\ell(\boldsymbol{\theta}) = \log p(\mathbf{y} \mid \boldsymbol{\theta})$ be the log–likelihood and denote with $\ell^{(3)}_{\hat{\boldsymbol{\theta}}} \in \mathbb{R}^{p \times p \times p}$ its third–order derivative evaluated at the maximum a posteriori (MAP) $\hat{\boldsymbol{\theta}}$. Then, a tractable skew–modal approximation for the generic posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ has density defined as

$$\hat{p}_{\text{SKS}}(\boldsymbol{\theta}) = 2\phi_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}; \mathbf{J}^{-1}_{\hat{\boldsymbol{\theta}}})\Phi[(\sqrt{2\pi}/12)\ell^{(3)}_{\hat{\boldsymbol{\theta}},stl}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_s(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_t(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_l], \qquad (3)$$

where $\mathbf{J}^{-1}_{\hat{\boldsymbol{\theta}}}$ denotes the inverse of the observed information matrix, evaluated at $\hat{\boldsymbol{\theta}}$, whereas $\ell^{(3)}_{\hat{\boldsymbol{\theta}},stl}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_s(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_t(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})_l$ is a cubic function of the parameters in $\boldsymbol{\theta}$, defined through index notation and with the Einstein's summation convention. As clarified in [16], (3) is a special case of the SKS density within (2) (with $\mathbf{z} = \boldsymbol{\theta}$), which interestingly coincides with a simple skewed perturbation of the original Laplace approximation $\phi_p(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}; \mathbf{J}^{-1}_{\hat{\boldsymbol{\theta}}})$. This connection has two important implications. First, from an inference perspective, the newly–derived skew–modal approximation is essentially as tractable as the Gaussian from the Laplace method, both in terms of density evaluation and i.i.d. sampling. Second, as show in [16], it is possible to prove that, under standard assumptions, the total variation distance among such a skewed approximation and the target posterior has a rate of convergence which improves by a $\sqrt{n}$ factor the one proved under the Bernstein–von Mises theorem for the Gaussian approximation from the Laplace method.

The above results open several stimulating directions of future research. Important examples include (i) the derivation of broadly–applicable skewed perturbations for general symmetric approximations from, e.g., variational Bayes and expectation–propagation, (ii) the theoretical study of high–dimensional regimes, and (iii) extensions to valid approximating densities of order higher than the third.

Conclusion and Acknowledgments. Skew–normals and related families, pioneered by the work of Professor Adelchi Azzalini, play a key role also in Bayesian inference. Nonetheless, such a role has been partially overlooked to date. The results summarized in this short article open important avenues for improved posterior inference via novel closed-form expressions, effective Monte Carlo methods, and more accurate approximations with theoretical support. Strengthening the collaboration among Bayesian statisticians and experts in the field of skewed distributions will yield impactful future advancements along these directions.

Among the many outstanding researchers with whom I had the luck to work on these topics over the past years, I would like to thank, in particular, my former Ph.D. students Niccolò Anceschi, Augusto Fasano and Francesco Pozza. Working with them has been a wonderful and enjoyable experience. Several of the results summarized in this short article would have been never obtained without their precious ideas, dedication and creativity.

References

1. Anceschi, N., Fasano, A., Durante, D., Zanella, G.: Bayesian conjugacy in probit, tobit, multinomial probit and extensions: a review and new results. J. Am. Stat. Assoc. **118**, 1451–1469 (2023)
2. Arellano-Valle, R.B., Azzalini, A.: On the unification of families of skew-normal distributions. Scand. J. Stat. **33**, 561–574 (2006)
3. Arellano-Valle, R.B., Branco, M.D., Genton, M.G.: A unified view on skewed distributions arising from selections. Can. J. Stat. **34**, 581–601 (2006)
4. Arellano-Valle, R.B., Genton, M.G.: Multivariate unified skew-elliptical distributions. Chil. J. Stat. **1**, 17–33 (2010)
5. Arellano-Valle, R.B., Genton, M.G., Loschi, R.H.: Shape mixtures of multivariate skew-normal distributions. J. Multivar. Anal. **100**, 91–101 (2009)
6. Arnold, B.C., Beaver, R.J.: Hidden truncation models. Sankhyā A 23–35 (2000)
7. Arnold, B.C., et al.: Skewed multivariate models related to hidden truncation and/or selective reporting. TEST **11**, 7–54 (2002)
8. Azzalini, A.: A class of distributions which includes the Normal ones. Scand. J. Stat. **12**, 171–178 (1985)
9. Azzalini, A., Capitanio, A.: Statistical applications of the multivariate skew normal distribution. J. Roy. Stat. Soc. B **61**, 579–602 (1999)
10. Azzalini, A., Capitanio, A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. Roy. Stat. Soc. B **65**, 367–389 (2003)
11. Azzalini, A., Capitanio, A.: The Skew-Normal and Related Families. Cambridge University Press, Cambridge (2014)
12. Azzalini, A., Dalla Valle, A.: The multivariate skew-normal distribution. Biometrika **83**, 715–726 (1996)
13. Branco, M., Genton, M.G., Liseo, B.: Objective Bayesian analysis of skew-t distributions. Scand. J. Stat. **40**, 63–85 (2013)
14. Canale, A., Kenne Pagui, E.C., Scarpa, B.: Bayesian modeling of university first-year students' grades after placement test. J. Appl. Stat. **43**, 3015–3029 (2016)
15. Durante, D.: Conjugate Bayes for probit regression via unified skew-normal distributions. Biometrika **106**, 765–779 (2019)
16. Durante, D., Pozza, F., Szabo, B.: Skewed Bernstein-von Mises theorem and skew-modal approximations (2024). arXiv:2301.03038
17. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman & Hall, London (1990)
18. Fasano, A., Durante, D.: A class of conjugate priors for multinomial probit models which includes the multivariate normal one. J. Mach. Learn. Res. **23**, 1358–1383 (2022)
19. Gupta, A.K., Gonzalez-Farias, G., Dominguez-Molina, J.A.: A multivariate skew normal distribution. J. Multivar. Anal. **89**, 181–190 (2004)

20. Karling, M.J., Durante, D., Genton, M.G.: Conjugacy properties of multivariate unified skew-elliptical distributions. arXiv:2402.09837 (2024)
21. Liseo, B., Loperfido, N.: A Bayesian interpretation of the multivariate skew-normal distribution. Stat. Probab. Lett. **61**, 395–401 (2003)
22. Liseo, B., Loperfido, N.: A note on reference priors for the scalar skew-normal distribution. J. Stat. Plan. Inference **136**, 373–389 (2006)
23. Onorati, P., Liseo, B.: An extension of the unified skew-normal family of distributions and application to Bayesian binary regression. arXiv:2209.03474 (2022)
24. Wang, J., Boyer, J., Genton, M.G.: A skew-symmetric representation of multivariate distributions. Stat. Sin. **14**, 1259–1270 (2004)

Statistical Learning in Healthcare: Towards a New Paradygm of Research

Francesca Ieva[1,2(✉)]

[1] Politecnico di Milano, Milan 20133, Italy
francesca.ieva@polimi.it
[2] Human Technopole, Health Data Science Center, Milan 20154, Italy
https://sites.google.com/view/francesca-ieva/home

Abstract. The availability of a vast amount of complex data coming from new technologies is revolutionizing healthcare research, providing a huge potential to both preventive and prognostic activities. To properly synthetize and manage the information deriving from different sources of clinical data (e.g., texts, medical imaging, omics), advanced methods are required, shifting the priority from predictive modeling to representation learning and fingerprint extraction.

In this paper, we will discuss the challenges that this new perspective introduces in healthcare research, focusing on how health analytics is able to provide meaningful answers.

Keywords: Health Analytics · Representation Learning · Healthcare Research · Fingerprint extraction

1 Background and Motivations

Health analytics is the process of analyzing healthcare data to make predictions on clinical endpoints, improve medical research, and better manage the offer as well as the quality of care. The field covers a broad range of disciplines, from epidemiology to (bio)statistics, from computer science to health economics, and offers insights on both the macro and micro level of the healthcare process. Health analytics is a data-centric discipline, aimed at enhancing an evidence based decision making both in clinical practive and in healthcare management.

The role of data in such a perspective is crucial, thus implying the need of a new paradygm of research, starting from data engineering to methods development and results communication. In fact, despite the opportunities multiple data sources provide for a structural health monitoring and for a more comprehensive description of patients' status, the aggregation of information from these sources remains a challenge. In the last decades, so much has been written on the benefits of (big) data for healthcare in improving patient outcomes, public health surveillance, and healthcare policy decisions [1]. However, this promise was often broken, due to multiple reasons. In particular, if on the one hand data are re-shaping the contemporary healthcare system with significant progress to

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 10–15, 2025.
https://doi.org/10.1007/978-3-031-64346-0_2

clinical investigation, on the other hand the challenges related to management, accessibility and standardized usage of data have sometimes made such progresses less impactful or even useless [2]. Moreover, the complexity of the data often requires methodological innovation the clinical community is not ready to receive [3]. Finally, accessibility is still an issue, hampering data integration and their comprehensive treatment.

For all these reasons, there is a crucial necessity to define standardized pipelines for accessing, treating and merging different data sources to make them available for predictive purposes, as well as advanced methodologies to extract knowledge from data and let them support decision making. The way decision making is carried out in healthcare is twofold: *Personalized Medicine* and *Precision Policies*. In the former, treatment planning and interventions of a given risk category are tuned according to the specific features each individual does show. In the latter, population data and related scenarios allow to optimize protocols and pathways of care to offer a better service to the patients. They are in some ways complementary perspectives, both essential for healthcare research. In both cases, new models and advanced statistical and computational methods are needed to properly tackle complexities and to develop robust and reliable inference.

In the following, we will try to envision a way for pursuing the goal of a comprehensive system of health analytics for complex healthcare data.

2 Challenges of Clinical Data

Clinical Data are characterized by many layers of complexity. Such multi-layer complexity is firstly related to the process of extracting a usable form of the information from raw data (*within source complexity*), then there is a second layer, related to the integration of different sources of information (*between source complexity*) into a single fingerprint or a comprehensive model aimed at scoring, risk-stratifying or predicting any relevant endpoint of interest. Both layers require methodological and technological innovation the clinical community is not always ready to receive.

In the last decade, despite remaining the prediction one of the main task of the analysis of clinical data, the availability of wider and more complex sources of information shifted the focus and the challenge of health analytics from the set up of a predictive model to methods for knowledge synthesis and extraction. Typical examples come from Natural Language Process (NLP) techniques for textual embeddings, radiomics for images and Machine Learning based representation methods for both multi-omics and more traditional, but large, databases. These data present intrinsic complexities and challenges in the way the relevant information can be captured, then plugged into models which enable risk scoring, stratification and prediction. Despite the broad nature of such data, the common requirement they advocate for is the construction of suitable (possibly low dimensional) embeddings where its information content may be enclosed.

Textual Data. The emergence of Electronic Health Records (EHR) has led to a vast amount of clinical data automatically stored in healthcare providers' data warehouses. However, a substantial portion of this data exists only in an unstructured format, comprising documents such as clinical notes, reports, discharge letters and referrals [4]. While these textual records harbour valuable information, their exploitation requires the application of NLP techniques. Applying NLP to healthcare documents introduces unique challenges compared to general-purpose texts. The specialised lexicon within medical contexts is characterised by a heightened level of ambiguity and an abundance of context-specific abbreviations. Furthermore, these documents often exhibit spelling errors, syntactical variations deviating from spoken language norms, and a lack of standardised structure, with considerable disparities observed across different healthcare providers, including clinicians and hospitals [5].

This underscores the necessity to develop tailored models specifically designed for the Italian language, given that existing literature.

Medical Imaging. Medical imaging is a second crucial asset in healthcare research. Among others, oncology is the field that benefits most of the analysis of images for supporting clinical decision making and personalized treatments. In fact, over the last two decades, the texture analysis of digital images arose as a valuable non-invasive proxy for biological assessment of tumors, eventually growing into a discipline of its own, named radiomics [6]. Expressly, the macroscopic appearance of tumors has been acknowledged as a valuable tool to guide clinical decisions in defining disease severity and treatment planning. Broadly speaking, image texture analysis consists of extracting descriptors of spatial variation of voxel grey scale and intensity within the image Volumes Of Interest (VOI), i.e., the tumor lesions. Under the name of radiomic features, such textural descriptors form a high dimensional vector embedding of the VOI and may provide a non-invasive assessment of tumor appearance from routinely acquired imaging studies [7]. These features are supposed to supply additional predictive and prognostic information, ready to postulate the underlying biological mechanisms of disease progression [8].

The dimensionality of the radiomics, together with their intrinsic redundancy, call for dimensionality reduction methods to be applied to the extraction of suitable fingerprint able to account for complexity of the imaging information, e.g., the heterogeneity deriving from multiple lesions captured over multiple time istances for the same patient.

Multi-omics Data. The concept and advancements of *Personalized Medicine* have been tightly coupled with our understanding of the genetic basis of health and disease, and its role in determining a large part the mentioned variability between individuals [9]. This is known as *omics revolution* [10]. In particular, the fields of genetics - the study of genes and their roles in inheritance - and genomics - the study of a person's genome and the interactions within the genome and the

greater environment - have fuelled personalized medicine thanks to new technologies allowing for cheaper and faster genetic testing and sequencing [11]. Moreover, nowadays health and disease are increasingly understood to be determined by many factors besides genes, such as transcriptomics, proteomics, metabolomics, microbiomics - all falling under the aforementioned -omics umbrella -, but also clinical and personal characteristics, lifestyle and environmental determinants and modifiers.

Including the omics information into predictive models and/or risk scoring algorithms means allowing for personalized treatments as well as for causality assessment.

Time-Varying Information. The role of epidemiological models is crucial for informing public health. However, traditional epidemiological models often neglect to consider the time-varying nature of exposure, treatment effect, behavioral variables, etc., thus introducing biases in the estimation of model parameters. Indeed, it is more and more often the case that the dynamic of a given process is more informative for the clinical practice and/or the clinical decision making than the actual value of the variable itself. With technological advances, intensive longitudinal data are increasingly generated by studies that repeatedly administer assessments over time. This offers unique opportunities to describe temporal behavioral changes in detail. Traditional analytical approaches impose strong parametric assumptions about the nature of change in the relationship between time-varying covariates and outcomes of interest.

The main challenge offered by including time-varying information into the models is to explicitly model changes in the association between exposure and outcome in a flexible manner.

3 A New Paradygm of Research in Health Analytics

Possible methods for fingerprint extraction and representation of information content of complex data like those mentioned in Sect. 2 can be found in [12–17]. Leveraging on these solutions, the future perspective would be to enclose them all in a flexible modeling environment to allow a fingerprint extraction, based on the joint use of all these sources together, thus allowing for a "molecular to individual" bridge to happen. Moreover, systems allowing the continuous update of predictions - like Digital Twins [18] - are essential for incorporating the dynamic information carried out by the data. Finally, inserting such tools in a microsimulation environment [19] would enable enhance scenario and sensitivity analyses to identify optimal practices and/or policies on-the-fly.

In the new perspective we would like to foster, any system aimed at supporting clinical decision making and at enhancing personalized prevention/treatment should rely on the following pillars: (a) the innovative methodological environment for fingerprint extraction from complex and diverse healthcare data (Fig. 1, left panel), (b) the comprehensive approach to healthcare data integration and

the advanced multi-modal fusion techniques, allowing a dynamical and adaptive inclusion of such data into predictive models (Fig. 1, right panel), (c) the availability of multiple enriched sources of observational data within large clinical and administrative registries and biobanks, and (d) the investigation of how each data source affects the prediction of diseases.

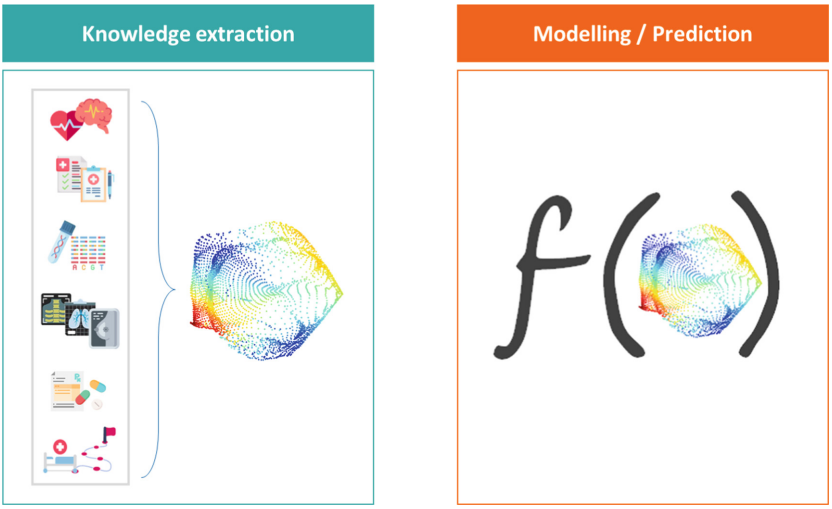


Fig. 1. Sketch of health analytics approach to a) fingerprint extraction from complex data (left panel) and b) predictive modeling (right panel).

This can be achieved by developing innovative systems able to evolve with the patient health, like a virtual twin following the individual in her/his health journey and recording health conditions from all possible sources of health data. Despite this is far from becoming a reality in the next future, models which focus on holistic descriptions of individuals' health status, and the introduction of comprehensive tools for diseases risk estimation, are definitively the first essential steps to be pursued, together with the set up of proper rules for data access and federated learning.

Acknowledgments. The author acknowledges the support by MUR, grant Dipartimento di Eccellenza 2023–2027.

The results mentioned in the paper were possible thanks to the effort of a number of people from both the Health Analytics group at MOX - Modeling and Scientific Computing lab, within the Department of Mathematics of Politecnico di Milano and the Health Data Science Center of Human Technopole.

References

1. Househ, M.S., Aldosari, B., Alanazi, A., Kushniruk, A.W., Borycki, E.M.: Big Data, Big Problems: a Healthcare Perspective. Stud Health Technol. Inform. **238**, 36–39 (2017). PMID: 28679881
2. Assidi, M., Buhmeida, A., Budowle, B.: Medicine and health of 21st Century: not just a high biotech-driven solution. NPJ Genom. Med. **7**(1), 67 (2022). https://doi.org/10.1038/s41525-022-00336-7.PMID: 36379953; PMCID: PMC9666643
3. https://www.politico.com/news/2022/08/15/artificial-intelligence-health-care-00051828
4. Tayefi, M., et al.: Challenges and opportunities beyond structured data in analysis of electronic health records, p. e1549. Computational Statistics, Wiley Interdisciplinary Reviews (2021)
5. Iroju, O.G., Olaleke, J.O.: A systematic review of natural language processing in healthcare. Int. J Inf. Techn. Comput. Sci. **8**, 44–50 (2015)
6. Mayerhoefer, M.E., et al.: Introduction to radiomics. J. Nucl. Med. **61**(4), 488–495 (2020)
7. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)
8. Chicklore, S., Goh, V., Siddique, V., Roy, A., Marsden, P.K., Cook, G.: Quantifying tumour heterogeneity in 18 f-fdg pet/ct imaging by texture analysis. Eur. J. Nucl. Med. Mol. Image **40**(1), 133–140 (2013)
9. MacEachern, S.J., Forkert, N.D.: Machine learning for precision medicine. Genome **64**(4), 416–425 (2021)
10. Collins, F.S., Varmus, H.: A new initiative on precision medicine. N. Engl. J. Med. **372**(9), 793–795 (2015)
11. Wu, P.Y., Cheng, C.W., Kaddi, C.D., Venugopalan, J., Hoffman, R., Wang, M.D.: Omic and electronic health record big data analytics for precision medicine. IEEE Trans. Biomed. Eng. **64**(2), 263–273 (2016)
12. Torri, V., Ercolanoni, M., Bortolan, F., Leoni, O., Ieva, F.: Clustering Italian medical texts: a case study on referrals. In: Proceedings of the Statistics and Data Science Conference. Pavia University Press, 2023
13. Savaré, L., Ieva, F., Corrao, G., Lora, A.: Capturing the variety of clinical pathways in patients with schizophrenic disorders through state sequences analysis. BMC Med. Res. Meth. **23**, 174 (2023). https://doi.org/10.1186/s12874-023-01993-7
14. Cavinato, L., Pegoraro, M., Ragni, A., Sollini, M., Erba, P.A., Ieva, F.: Imaging-based representation and stratification of intra-tumor heterogeneity via tree-edit distance. Sci. Rep. **12**, 19607 (2022)
15. Massi, M.C., et al.: Learning high-order interactions for polygenic risk prediction. PLoS ONE **18**(2), e0281618 (2023)
16. Franco, N.R., Massi, M.C., Ieva, F., et al.: Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. Radiother. Oncol. **159**, 241–248 (2021)
17. Spreafico, M., Ieva, F.: Functional modelling of recurrent events on time-to-event processes. Biom. J. **63**(5), 948–967 (2021)
18. Attaran, M., Celik, B.G.: Digital twin: benefits, use cases, challenges, and opportunities. Decis. Anal. J. **6**, 100165 (2023). https://doi.org/10.1016/j.dajour2023.100165
19. Çağlayan, Ç., Terawaki, H., Chen, Q., Rai, A., Ayer, T., Flowers, C.R.: Microsimulation modeling in oncology. JCO Clin. Cancer Inform. **2**, 1–11 (2018). https://doi.org/10.1200/CCI.17.00029.PMID: 30652551; PMCID: PMC6386553

Realistically Complex Spatio-Temporal Modelling – Responsibilities, Challenges and Fallacies

Janine B. Illian(✉)

University of Glasgow, Glasgow, Scotland, UK
janine.illian@glasgow.ac.uk
https://www.gla.ac.uk/schools/mathematicsstatistics/staff/janineillian/

Abstract. Statisticians are able to analyse data that are relevant to ecological and environmental research and that may help society address these challenges by developing increasingly complex methodology and models – as new technologies and new challenges arise. In this talk, we will explain that as statisticians we should not only see these societal challenges as opportunities for us to develop new methodology and publish papers in statistical journals, but also be aware of the responsibility we have towards society to use our skills to help resolve these issues.

Keywords: spatio-temporal modelling · spatial point processes · computational efficiency · statistical ecology

1 Introduction

These days, society is faced with many pressing issues often resulting from anthropogenic impact on the natural world, including climate change, biodiversity loss, environmental pollution and human and animal health issues resulting from the latter. To address these issues, ecological and environmental research aims to improve our understanding of how these complex issues arise and how they can be mediated to ultimately allow live on earth to continue. As a result, and through the availability of modern and continuously evolving technology, large amounts of increasingly detailed data are being collected, which have the potential to provide many new insights. At the same time society has understood the importance of statistics and data science, and has become increasingly aware of the important role that statistical data analyses have in dealing with issues facing modern societies.

As statisticians we are able to analyse data that are relevant to ecological and environmental research and that may help society address these challenges by developing increasingly complex methodology and models – as new technologies and new challenges arise. This was, for instance, clearly evident during the recent COVID-19 crisis, which created a huge number of statistical publications analysing the COVID-19 data – and even made ordinary citizens became "hobby statisticians".

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 16–20, 2025.
https://doi.org/10.1007/978-3-031-64346-0_3

In this talk, we will explain that as statisticians we should not only see these societal challenges as opportunities for us to develop new methodology and publish papers in statistical journals, but also be aware of the responsibility we have towards society to use our skills to help resolve these issues. We do have a responsibility to society to aid the analysis of data, and to help doing this with the right methodology that is accessible even to non-specialists. This responsibility is a broad responsibility and we will illustrate how it spans a number of activities that statisticians are involved in when developing statistical methodology, primarily focusing on

- the **relevance** of the methodology, i.e. the need to develop methods that are relevant in practice and hence enable others to answer relevant scientific problems;
- the **accessibility** of methodology and the software, i.e. the importance of making the methodology accessible through usable software to enable non-experts to use these and appropriately interpret the outcome from a modelling exercise;
- the **user-perspective** of disseminating and teaching the methodology, i.e. the challenge of taking approaches to teaching methodology with the users' perspective and aims in mind.

In this talk, we use the example of spatial point processes to illustrate specifically these different elements of our responsibility. Spatial point processes are models that model the pattern formed by objects or events in space. Figure 1 shows an example of such a pattern – the spatial pattern formed by the locations of more than 800 trees from the species *Leea aculeata* in a rainforest study plot in Danum, Malaysia, one of almost 700 tree species in the study plot that is of interest due to its outstanding biodiversity and the need for ecologists to understand the mechanisms that sustain this diversity, for example through understanding habitat preferences of different rainforest tree species [4][1].

For the discussion here we focus on a specific class of spatial point processes termed *log Gaussian Cox processes* These are well-studied and flexible point process models that are generalisations of inhomogeneous Poisson process with a random intensity function. Specifically, an inhomogeneous Poisson process N, is a point process, in which the number of points within a region $A \subset \Omega$, where $\Omega \subset \mathbb{R}^2$ is a bounded region, follows a Poisson distribution with mean $\Lambda(A) = \int_A \lambda(s)ds$, where $\lambda(s)$ is the intensity surface of the point process that describes the number of points per unit area. The likelihood of an inhomogeneous Poisson process is

[1] "The Danum plot is a core project of the Southeast Asia Rain Forest Research Partnership (SEARRP). We thank SEARRP partners, especially Yayasan Sabah for their support, and HSBC Malaysia and the University of Zurich for funding. We are grateful to the research assistants who are conducting the census, in particular the team leader Alex Karolus, and to Mike Bernados and Bill McDonald for species identifications. We thank Stuart Davies and Shameema Esufali for advice and training.".

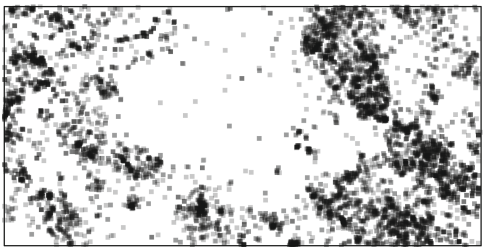


Fig. 1. Spatial pattern formed by the locations of trees of the rainforest tree species *Leea aculeata* in a 50 hektar plot in Danum Valley, Malaysia.

$$\pi(N|\lambda) = \exp\left\{|\Omega| - \int_{\Omega} \lambda(s)ds\right\} \prod_{s_i \in N} \lambda(s_i). \tag{1}$$

When $\lambda(s) = \log(Z(s))$ with $Z(s)$ a Gaussian random field, this process is often referred to as a log Gaussian Cox process. The construction based on the random field makes the process particularly flexible as it accounts for unexplained spatial structures – e.g. caused by dispersal limitation – and the class of models is hence versatile in its use for ecological data [2,6,9]. While the model and its likelihood look pretty straight forward at first sight, it turns out that this likelihood in itself is complicated. In the first place, it is analytically intractable, as the integral of the intensity function cannot be calculated explicitly and needs to be calculated using numerical methods. This makes fitting even relatively simple versions of this point process model computationally costly. However, this is only one of the complexities that are common in a spatial point process analysis and its use in concrete scientific data analyses.

In practice, the models fitted to a dataset need to be realistically complex to reflect natural processes and the questions of interest – and hence will be even more computationally costly. In other words, a model that is relevant in practice will perhaps have to be a marked point process that takes into account properties of the individuals represented by the points, e.g. older (and hence larger) trees might form different patterns than younger trees and hence an individual's size (typically expressed as diameter at breast height), would need to be taken into account and modelled along with the species location. Similarly, in other contexts, spatio-temporal processes might be relevant, where spatial structures vary with time. These realistically complex models make modelling increasingly computationally complex, which has lead to the development of computationally efficient approaches, in particular based on integrated nested Laplace approximation (INLA) [7,8,10,11].

However, developing methods that are complex enough to reflect natural processes appropriately – rather than being merely mathematically interesting and pleasing – is only one aspect of appreciating our responsibility for addressing the need for statistical support to society. The pattern in Fig. 1 shows a pattern where ecologists have been able to collect data on every tree above a certain

height within the observation area. However, in many ecological projects this is not feasible, since the observation area might be too big to be sampled in its entirety – it might be an entire ocean – or since the individuals of interest are not always detectable – such as diving sea mammals. In other words, the observation process that has been used to collect the data is not necessarily uniform in space and an individual's detectability might not be known – i.e. the observation process will have to be taken into account and modelled [5]. Specific methodology has been developed and made available through the software package `inlabru`, which addresses this issue and allows to fit complex statistical models, including spatial point process models, along with a number of other observation processes such as distance sampling and plot sampling based on the INLA approach, but with added features to increase practical relevance [1,13,14].

However, is this really everything we need to do to fulfil our role as applied statisticians – do we merely need to make methods available that have the potential to be used to address relevant scientific question? In this talk we will demonstrate that our responsibility does not end here. As briefly mentioned above, we also need to make sure that the methodology we develop can be used and interpreted by users who are non-experts. Consider the log Gaussian Cox process in Eq. (1) again. Even the simplest of these models, e.g. a simple model just containing a simple Gaussian random field and some spatial covariates is difficult to explain to a non-specialist as a stochastic process where the locations are being modelled and hence random. In addition, we are dealing with a doubly-stochastic point process here – a point process that varies relative to a Gaussian random field [3]. While even statisticians still struggle to devise appropriate ways of providing prior choices in the context [12], the concept of random fields has proved particularly difficult for us to explain to non-specialists as these mathematical constructs are not intuitively easy to understand. We will discuss how to best address this issue in particular, by arguing that it is not always helpful to explain a concept through its mathematical definition or characterisation. It is important to rather take on the users' perspective – why do they use the model and how do they want to interpret the output? It is often much more useful to discuss a concept, such as the random field here, through its *role* in the model and in the modelling process. In the context of log Gaussian Cox processes, the random field represents spatial structure that cannot be explained by covariates – and hence can be introduced as a model component whose role it is to reflect these unexplained structures.

Similarly, at the same time as providing software that has implemented our practically relevant methodology, we also have the responsibility to make the software accessible to users. This may be done through teaching or through publications such as text books – and again needs to take on the users' perspective. How can we structure the software so that it becomes more intuitive to use and that it best reflects the roles of the model components to the user? Finally, fitting a complex model, such as a, e.g., spatio-temporal log Gaussian Cox process, involves far more than writing the code that the specific `R`-library calls for. During the modelling process a number of difficult decisions have to be

made. These are often more daunting to the users than writing the actual code, and include, but are not limited to, decisions on prior choices when modelling is done in a Bayesian context, on the spatial resolution of the approximation of continuous space and on choices of model complexity. We will illustrate our current approaches to addressing these issues and use concrete examples taken from applications, teaching material and software implementations, in particular in the context of ecological modelling and the software package `inlabru`.

References

1. Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B.: Inlabru: an (R) package for (B)ayesian spatial modelling from ecological survey data. Methods Ecol. Evol. **10**(6), 760–766 (2019)
2. Baddeley, A., Rubak, E., Turner, R.: Spatial Point Patterns: Methodology and Applications with R. CRC Press, Boca Raton (2015)
3. Cox, D.R.: Some statistical methods connected with series of events. J. Roy. Stat. Soc.: Ser. B (Methodol.) **17**(2), 129–157 (1955)
4. Davies, S.J., et al.: Forestgeo: understanding forest diversity and dynamics through a global observatory network. Biol. Cons. **253**, 108907 (2021)
5. Hedley, S.L., Buckland, S.T.: Spatial models for line transect sampling. J. Agric. Biol. Environ. Stat. **9**(2), 181–199 (2004)
6. Illian, J., Penttinen, A., Stoyan, H., Stoyan, D.: Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, Hoboken (2008)
7. Illian, J.B., et al.: Fitting complex ecological point process models with integrated nested Laplace approximation. Methods Ecol. Evol. **4**(4), 305–315 (2013)
8. Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **73**(4), 423–498 (2011)
9. Møller, J., Waagepetersen, R.P.: Modern statistics for spatial point processes. Scand. J. Stat. **34**(4), 643–684 (2007)
10. Rue, H., Martino, S., Chopin, N.: Approximate bayesian inference for latent gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **71**(2), 319–392 (2009)
11. Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H.: Going off grid: computationally efficient inference for log-Gaussian Cox processes. Biometrika **103**(1), 49–70 (2016)
12. Sørbye, S.H., Illian, J.B., Simpson, D.P., Burslem, D., Rue, H.: Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. J. R. Stat. Soc.: Ser. C: Appl. Stat. **68**(3), 543–564 (2019)
13. Williamson, L.D., et al.: Comparing distribution of harbour porpoise using generalized additive models and hierarchical Bayesian models with integrated nested Laplace approximation. Ecol. Model. **470**, 110011 (2022)
14. Yuan, Y., et al.: Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. Ann. Appl. Stat. **11**(4), 2270–2297 (2017)

Specialized Sessions

Supervised Classification of Functional Data via Ensembles of Different Functional Representations

Donato Riccio[1], Fabrizio Maturo[2(✉)], and Elvira Romano[3]

[1] Department of Engineering and Science, Universitas Mercatorum, Rome, Italy
donato.riccio@studenti.unimercatorum.it
[2] Faculty of Technological and Innovation Sciences, Universitas Mercatorum, Rome, Italy
fabrizio.maturo@unimercatorum.it
[3] Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy
elvira.romano@unicampania.it

Abstract. Functional Data Analysis (FDA) has become popular in the statistical literature for modelling high-dimensional time series. Although supervised learning has been broadly explored from various perspectives, ensembles of functional classifiers have only lately emerged as a matter of substantial interest. The latter topic offers novel aspects and challenges from mixed statistical viewpoints. This article focuses on ensemble learning for functional data and offers a possible approach where distinct functional representations can be adopted to train ensemble members, and base-model predictions can be combined to improve classifiers' performances.

Keywords: FDA · supervised classification · statistical learning · diversity

1 Introduction

Functional Data Analysis (FDA) is a statistical technique that deals with situations where functions rather than scalar values represent observations. This approach involves modelling the underlying function that generates the data directly instead of the sequence of observations. Thus, the observed functional data is treated as single entities [1–4,6]. Over the past few decades, extensive research has delved into supervised classification challenges within the FDA literature. However, the intrinsic high dimensionality in functional data gives rise to the curse of dimensionality. The latter issue entails several damaging effects, such as data sparsity, model selection, multicollinearity, and distance concentration. The complexity of these challenges has highlighted the importance of robust statistical approaches for effectively analyzing and understanding such data.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 23–27, 2025.
https://doi.org/10.1007/978-3-031-64346-0_4

A key advantage of FDA is that it reduces the complexity of high-dimensional data. It is beneficial for classifying biomedical and other types of time series data without imposing constraints [5,9–11], In addition to the works that drew the attention of the international scientific community to FDA [1–3], many methodological proposals have been suggested to expand traditional classification techniques to the context of FDA. Recently, in the non-functional context, tree ensemble methods [7,8] have become famous for classification and regression problems. The main idea is to use multiple base learners and combine their predictions to improve accuracy and stability over using a single estimator. Therefore there is a rising interest in using ensemble methods also for functional data. Nevertheless, there are still noteworthy gaps in this area. More elaborate ensemble architectures, such as voting ensembles, must be explored further for functional data. Some crucial questions still need to be answered, such as defining criteria for selecting ensemble members and aggregate forecasts from base learners. For example, diversity is especially vital in the context of ensemble methods [12–15]. In the context of statistical learning, the latter refers to the level of disagreement among individual learners in an ensemble. Procedures such as bagging and boosting intentionally induce diversity among base learners in typical machine learning ensembles. Diverse ensembles can achieve lower generalization errors than individual learners by permitting learners to complete each other's deficiencies. However, ensembles with significantly low accuracy members are unlikely to enhance predictive performance, regardless of their diversity. There are still some open questions when creating ensembles of functional data. One of these questions is ensuring the ensemble members are diverse.

Our research aims to address some of the gaps in FDA classification problems and furnish instruments for operating ensemble techniques. One potential solution to the above-mentioned issue is to use different basis function representations. Hence, we have developed an ensemble architecture designed specifically for functional data. By aggregating diverse base learners, the ensemble reduces overfitting and may improve performances.

2 A Brief Introduction to Functional Data Analysis

Consider a set of N observations $\{(t_i, y_i)\}_{i=1}^{N}$ where t_i is a feature vector and y_i is the corresponding response value. We aim to estimate an underlying function, $x(t) : T \to \mathbb{R}$, that generated these observations. We assume that the function $x(t)$ can be defined as $x_i(t) = f(t_i) + \epsilon_i$, where $f(t_i)$ represents the true (but unknown) value of the function at time t_i, and ϵ_i represents the noise or error term at time t_i.

To turn the individual points into a functional representation, we need to find an approximation of the function $x(t)$ using an element from a function space. This function space is usually a Hilbert space H containing square-integrable functions on T, with an inner product defined by $\langle x(t), g(t) \rangle = \int_T x(t)g(t)\,dt$ with norm given by $\|x(t)\| = \sqrt{\langle x(t), x(t) \rangle} = \left(\int_T x^2(t)\,dt\right)^{1/2}$.

A popular method for representing functional data involves approximating $x(t)$ by a linear combination of basis functions. $\{\phi_k\}_{k=1}^{K}$ given by:

$$x(t) \approx \hat{x}(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{1}$$

For example, B-splines are a type of mathematical function that is very efficient and flexible when it comes to modeling various forms. They are smooth and continuous, meaning they don't have any sudden jumps or breaks in them. B-splines are defined as piecewise polynomial functions of degree o, which are smooth and continuous up to their $(o-1)^{th}$ derivative. The definition of a spline curve is given by $S(t) = \sum_{k=0}^{K} c_k B_{k,o}(t)$. Thus, the functional data can be represented as a linear combination of B-splines as follows:

$$x_i(t) \approx \sum_{k=1}^{K} c_k B_{k,o}(x) \tag{2}$$

Using B-splines in functional data analysis (FDA) transforms the task of analyzing functional data into a finite-dimensional problem. This makes it easier to apply machine learning techniques. B-splines are particularly advantageous in FDA because they can efficiently approximate complex functional forms while remaining flexible.

3 Supervised Learning of Functional Data via Ensembles of Different Functional Representations

Functional classification aims to predict an outcome Y by employing a predictor variable X taking values in a separable metric space (E, d). The present investigation is oriented towards scalar-on-function classification, where Y is a categorical binary variable. Thus, the method is intended for functional data of the form $\{x_i(t), y_i\}$, with a curve $x_i(t), t \in T$ as the predictor, and y_i as the response at sample $i = 1, ..., N$. Let Y take on the values 0 or 1. A mapping $f : F \to \{0, 1\}$, called a *binary classifier*, classifies a new observation $x_{new}(t)$ from X by mapping it to its predicted label.

Using B-splines, the features matrix is given by:

$$\mathbf{C} = \begin{bmatrix} c_{11} & \dots & c_{1K} \\ \vdots & \ddots & \vdots \\ c_{N1} & \dots & c_{NK} \end{bmatrix} \tag{3}$$

where c_{ik} is the coefficient for the ith curve $i = 1, ..., N$ relative to the kth $k = 1, ..., K$ basis function $\phi_k(t)$ in the linear combination.

Ensemble learning involves incorporating multiple models to acquire better predictive performance. The main idea is that diverse and independently trained learners can exceed any individual model. By aggregating predictions across a set of models, individual learners' strengths can be leveraged while their weaknesses

are overcome. Diversity in ensemble learning refers to the degree of disagreement between the individual learners. If an ensemble has greater diversity, it becomes more robust and can provide improved predictions compared to unique models. In FDA, diversity may occur when operating different base approximations. This can intrinsically lead models to capture divergent characteristics and increase disagreement. Let us define a training set $\{(x_i, y_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i = (x_{i1}, ..., x_{iP})$ is a P-dimensional vector of predictor variables for observation i, and $\boldsymbol{y}_i \in \{1, \ldots, Z\}$ is the categorical response variable with Z classes.

To obtain multiple functional representations, we first approximate each vector $\mathbf{x}_i$ as a function $x_i(t)$ using B-spline basis expansions of orders $o = 1, \ldots, O$:

$$x_i^{(o)}(t) = \sum_{j=1}^{K^{(o)}} c_{ij}^{(o)} B_j^{(o)}(t) \tag{4}$$

where $K^{(o)}$ is the number of B-spline bases used for order o.

Let $\mathbf{C}^{(o)} \in \mathbb{R}^{N \times K^{(o)}}$ be the B-spline coefficient matrix for order o, obtained by stacking the coefficient vectors $\mathbf{c}_i^{(o)}$ as rows. For each o, the scores $\mathbf{C}^{(o)}$ are adopted to train a model $f^{(o)} : \mathbb{R}^{K^{(o)}} \to \{1, \ldots, Z\}$, where any algorithm can be used (e.g., KNN, classification trees, and random forest). The result is an ensemble of O models $\{f^{(1)}, \ldots, f^{(O)}\}$, where each $f^{(o)}$ is trained on the functional representations from B-spline order o. For a new test function, we obtain representations $\mathbf{c}^{(o)}$ under each optimal B-spline basis. These are fed into the trained models $f^{(o)}$ to obtain predicted class labels $\hat{y}^{(o)} = f^{(o)}(\mathbf{c}^{(o)})$.

Let $\mathbf{F} \in \mathbb{R}^{M \times O}$ be the prediction matrix containing the predictions from the O models on the M test examples. The element $\mathbf{F}_{mo} = \hat{y}_n^{(o)}$ is the predicted class label from model o on test input $x_m(t)$:

$$\mathbf{F} = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \cdots & \hat{y}_1^{(O)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \cdots & \hat{y}_2^{(O)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_M^{(1)} & \hat{y}_M^{(2)} & \cdots & \hat{y}_M^{(O)}. \end{bmatrix} \tag{5}$$

Finally, the majority vote is used to select the final prediction as follows:

$$\hat{y}_m = \arg \max_{u \in 1, \ldots, Z} \sum_{o=1}^{O} \mathbb{I}(\mathbf{F}_{mo} = u) \tag{6}$$

where $\mathbb{I}(\mathbf{F}_{mo} = u)$ is an M-dimensional indicator vector that has a 1 for test examples where model o predicted class u, and 0 otherwise.

4 Conclusions

This study shows an ensemble strategy designed for supervised classification tasks that involve functional data. Instead of relying on raw multivariate observations, the suggested approach suggests representing functions approximated under different bases to leverage the information power of a diverse ensemble.

Nonetheless, there is a potential disadvantage to this method, which is that it may be less interpretable compared to a single model because of the multiple representations and voting aggregation. Nonetheless, model-agnostic variable importance measures could be used to extract insights. Additionally, training multiple models incurs a higher computational cost when compared to a single classifier.

References

1. Ferraty, F., Vieu, P.: Curves discrimination: a nonparametric functional approach. Comput. Stat. Data Anal. **44**(1–2), 161–173 (2003). https://doi.org/10.1016/s0167-9473(03)00032-x
2. Ramsay, J., Silverman, B.: Functional Data Analysis, 2nd edn. Springer, New York (2005). https://doi.org/10.1007/b98888
3. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer, Cham (2006). https://doi.org/10.1007/0-387-36620-2
4. Cuevas, A.: A partial overview of the theory of statistics with functional data. J. Stat. Plan. Inference **147**, 1–23 (2014). https://doi.org/10.1016/j.jspi.2013.04.002
5. Maturo, F., Verde, R.: Pooling random forest and functional data analysis for biomedical signals supervised classification: theory and application to electrocardiogram data. Stat. Med. **41**(12), 2247–2275 (2022). https://doi.org/10.1002/sim.9353
6. Ferraty, F.: Recent Advances in Functional Data Analysis and Related Topics. Physica-Verlag HD (2011). https://doi.org/10.1007/978-3-7908-2736-1
7. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2004). https://doi.org/10.1023/A:1010933404324
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
9. Yu, Y., Lambert, D.: Fitting trees to functional data, with an application to time-of-day patterns. J. Comput. Graph. Stat. **8**(4), 749–762 (1999). https://doi.org/10.1080/10618600.1999.10474847
10. Maturo, F., Verde, R.: Supervised classification of curves via a combined use of functional data analysis and tree-based methods. Comput. Stat. **38**(1), 419–459 (2023). https://doi.org/10.1007/s00180-022-01236-1
11. Maturo, F., Verde, R.: Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. Comput. Stat. **39**, 239–270 (2024). https://doi.org/10.1007/s00180-022-01236-1
12. Brown, G., Kuncheva, L.I.: "Good" and "Bad" diversity in majority vote ensembles. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 124–133. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12127-2_13
13. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996). https://doi.org/10.1007/bf00058655
14. Breiman, L., et al.: Classification and Regression Trees. Chapman and Hall/CRC (1984). https://doi.org/10.1201/9781315139470
15. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. **51**(2), 181 (2003). https://doi.org/10.1023/A:1022859003006

Conformal Based Uncertainty Bands for Predictions in Functional Ordinary Kriging

Anna De Magistris(✉), Andrea Diana, and Elvira Romano

University of Campania Luigi Vanvitelli, Caserta, Italy
anna.demagistris@unicampania.it

Abstract. Functional Ordinary Kriging is the most widley used method to predict a curve at a given spatial point by computing a weighted average of observed functional values in space. Determining the uncertainty associated with predicted curves remains a challenge. To address this issue we introduce a distribution free prediction method for the proposed model and illustrate its performances using the well known data set of Canadian Temperature [12].

Keywords: conformal prediction · ordinary kriging · spatio-functional data

1 Introduction

In spatial functional data analysis [1], a common aim is to predict a curve in given target location [5,10,11]. This prediction task involves incorporating both the functional aspect of the data and the spatial dependence or correlation structure to make informed predictions at these unobserved spatial locations. Functional Kriging is a method that estimates a weighted average of observed functional values, in the unobserved spatial location, with weights given to nearby observations with higher functional correlation. The simplest form of Kriging, known as Ordinary Kriging, assumes a constant mean value to predict curves at locations that are not monitored. The uncertainty associated with predicted curves is usually based on resampling methods such as bootstrap or permutation techniques. This allows confidence bands to be derived for functional predictions [6,9]. These methods take into account the inherent uncertainty of data and prediction processes. Nevertheless, research is ongoing to develop alternative approaches that directly address uncertainty in the functional context. One approach is the Conformal Prediction (CP) method [4]. In this paper, we investigate a new procedure introduced in [2,3] that involves the use of Conformal Prediction to establish a region of uncertainty for predictions obtained through functional Kriging techniques.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 28–33, 2025.
https://doi.org/10.1007/978-3-031-64346-0_5

2 Method

Consider a geostatistical functional stochastic process $\{X_s(t) : s = (u, v) \in D \subseteq \mathbb{R}^2\}$ whose functions $X_{s_i}(t)$ are random functions located in n points and for each couple of $s_i = (u_i, v_i)$, where u_i represents latitude and v_i is longitude. Each function is defined on $T = [a, b] \subseteq \mathbb{R}$ and is assumed to belong to a Hilbert space $L^2(T) = \{f : T \to \mathbb{R} \mid \int_T f(t)^2 dt < \infty\}$ with the inner product $\langle X_{s_i}, X_{s_j} \rangle = \int_T X_{s_i}(t) X_{s_j}(t) dt$ [12]. For a fixed site s_i, it is assumed that the observed functions can be expressed according to the model:

$$X_{s_i}(t) = \mu_{s_i}(t) + \epsilon_{s_i}(t), \quad i = 1, \ldots, n$$

where deterministic component $\mu_{s_i}(t)$ describes the non-constant spatial mean variation and a stationary stochastic component $\epsilon_{s_i}(t)$ is supposed to be a zero-mean. Furthermore, assume that the stochastic process is second-order stationary and isotropic, i.e.:

- $\mathbb{E}[X_{s_i}(t)] = \mu(t), \quad \forall i = 1, \ldots, n;$
- $\mathbb{V}[X_{s_i}] = \sigma^2(t), \quad \forall i = 1, \ldots, n;$
- $Cov(X_{s_i}(t), X_{s_j}(t)) = \mathbb{E}[\langle X_{s_i}(t), X_{s_j(t)} \rangle] = C(h), \quad \forall i, j = 1, \ldots, n, \quad h = ||s_i - s_j||.$

Given n observations $\{X_{s_1}(t), \ldots, X_{s_i}(t), \ldots, X_{s_n}(t)\}$ the aim is to formulate the Oridnary Kriging predictor for functional data to predict the variable X_{s_0} located in $s_0 \in D$. The best linear unbiased predictor (BLUP) [7] and [8] for $X^*_{s_0}$ is given by:

$$X^*_{s_0} = \sum_{i=1}^{n} \lambda^*_i X_{s_i}$$

whose weights $\lambda^*_1, \ldots, \lambda^*_n$ minimize the global variance of the prediction error under the unbiasedness constraint:

$$(\lambda^*_1, \ldots, \lambda^*_n) = \underset{\lambda_1, \ldots, \lambda_n \in \mathbb{R}}{\operatorname{argmin}} \mathbb{V}(X^*_{s_0} - X_{s_0}) \quad t.c. \quad \sum_{i=1}^{n} \lambda_i = 1.$$

where the condition $\sum_{i=1}^{n} \lambda_i = 1$ is the unbiasedness constraint.

To evaluate the uncertainty of a predicted curve $X^*_{s_0}(t)$ from a new site s_0 without making any assumptions about the distribution of the data we considered a Conformal Prediction (CP) method. Given nominal miscoverage level $\alpha \in (0, 1)$, we can define a prediction band $C \subset L^2(T) \times D$ based on the observed data $X_{s_1}(t), \ldots, X_{s_i}(t), \ldots, X_{s_n}(t)$ such that the probability of the true curve $X_{s_0}(t)$ falling within the band is at least $1 - \alpha$, as expressed by equation:

$$\mathbb{P}(X_{s_0} \in C(s_0)) \geq 1 - \alpha$$

where $C(s_0)$ represents the set of curves in $L^2(T)$ where $X_{s_0}(t)$ is contained, for a given point $s_0 \in D$. In the context of functional data analysis, Diquigiovanni

et al. in [5], introduced the concept of a prediction band. To facilitate notation we indicate the data with $\boldsymbol{z}_{s_i}(T) = (s_i, X_{s_i}(t))$ with $t \in T$. The main steps of the proposed Conformal Prediction procedure can be described by the following procedure:

- Definition of the set (Z_{TRAIN}) and of the set (Z_{TEST}).
  Indicated with Z_{TRAIN} the data sample on which to calibrate the ordinary kriging model and Z_{TEST} the set of data with which to construct the band, in accordance with what Waldo Tobler states: "everything is related to everything else, but nearby things are more related than distant ones", for the separation of the original data into the two sets Z_{TRAIN} and Z_{TEST} we consider the closest L sites in s_0, creating the set:

$$Vic(s_0) = \{\boldsymbol{z}_{s_i}(T) : d_{i,0} < \delta_L\}$$

  where δ_L is the chosen proximity threshold. Currently δ_L is the median value of the distribution of spatial distances $d_{i,0}$. So we define Z_{TRAIN} the M closest elements of $Vic(s_0)$, then:

$$Z_{TRAIN} = \{\boldsymbol{z}_{s_i}(T) \in Vic(s_0) : d_{i,j} < \delta_M\}, \quad Z_{TEST} = Z \setminus Z_{TRAIN}.$$

  The idea of conformal prediction is to try all possible curves for the test object to see how well these curves conform to the set of training examples.
- Estimation of the trace-variogram with errors for the data as a whole $TRAIN$ (Z_{TRAIN}).
  In order to perform kriging we estimate the semivariogram model $\gamma(h) = \frac{1}{2}\mathbb{V}(X_{s_i} - X_{s_j}) = \mathbb{E}[(X_{s_i} - X_{s_j})]^2$. The theoretical trace-variogram can be estimated as:

$$\hat{\gamma}(h) = \frac{1}{|2N(h)|} \sum_{(i,j)\in N(h)} ||X_{s_i} - X_{s_j}||^2$$

  where $N(h) = \{(i,j) \ : ||s_i - s_j|| = h\}$ and $|N(h)|$ indicates the cardinality.
- Estimation of the covariance and prediction $X^*_{s_0}$ with data $\boldsymbol{z}_{s_i}(T) \in Z_{TRAIN}$.
  Once the trace variogram has been estimated, we use the $\boldsymbol{z}_{s_i}(T) \in Z_{TRAIN}$ data set to predict $X^*_{s_0}$ in s_0 as explained above.
- Definition of the radius and modulation for prediction band for $X^*_{s_0}$starting from the data $\boldsymbol{z}_{s_i}(T) \in Z_{TEST}$.
  In order to create the prediction bands we use different combination of modulation functions and non-conformity measures. The following two modulation functions are used:

$$\mathcal{S}_{sup}(t) := sup_{\hat{X}_{s_{0k}} \in \tilde{X}} |X^*_{s_0}(t) - \hat{X}_{s_{0j}}(t)|; \tag{1}$$

$$S_{sqrt}(t) := \sqrt{\frac{\sum_{\hat{X}_{s_{0k}} \in \tilde{X}} (X^*_{s_0}(t) - \hat{X}_{s_{0j}}(t))^2}{|\tilde{X}|}}. \tag{2}$$

The following two non-conformity measures are used:

$$\mathcal{D}_{sup}\left(\frac{\hat{X}_{s_{0j}}(t)}{\mathcal{S}(t)};\frac{X^*_{s_0}(t)}{\mathcal{S}(t)}\right) := sup_{t\in T}\left|\frac{X^*_{s_0}(t)-\hat{X}_{s_{0j}}(t)}{\mathcal{S}(t)}\right| \tag{3}$$

$$D_{sqrt}\left(\frac{\hat{X}_{s_{0j}}(t)}{\mathcal{S}(t)};\frac{X^*_{s_0}(t)}{\mathcal{S}(t)}\right) := \sqrt{\int\frac{(X^*_{s_0}(t)-\hat{X}_{s_{0j}}(t))^2dt}{\mathcal{S}(t)}}. \tag{4}$$

The different proposed combinations are implemented by the following algorithm:

Algorithm 1. Conformal Prediction

Input: Z_{TRAIN} observations, Z_{TEST} observations, s_0 prediction position.
Output: $\rho^{\mathcal{S}}$ prediction band radius, modulation function $\mathcal{S}$.

```
X*_{s0} ← prediction with Ordinary Kriging on Z_TRAIN
X̃ ← construction of the set as follows
for z_{s_j} ∈ Z_TEST do
    Ẑ ← {Z_TRAIN, z_{s_j}}
    X̂_{s0j} ← prediction with Ordinary Kriging on Ẑ
    X̃_j ← X̂_{s0j}
end for
S(t) ← define the modulation function
for X̃_{s0j} ∈ X̃ do
    R_j ← D( X̂_{s0j}(t)/S(t) ; X*_{s0}(t)/S(t) ) non-conformity scores
end for
ρ^S ← (1 − α)-th percentile of distribution of R_j.
```

- Band prediction. We build the prediction band for $X^*_{s_0}$ with the defined radius and modulation function as follows:

$$C_{pred}(s_0) = \{f(t)\in L^2(T) : X^*_{s_0} - \rho^{\mathcal{S}}\mathcal{S}(t) \le f(t) \le X^*_{s_0} + \rho^{\mathcal{S}}\mathcal{S}(t)\ \forall t\in T\}$$

3 Results

To demonstrate the main results of the proposed method we consider the well-known Canadian temperature dataset [12]. The data set consists of daily annual mean temperature collected at 35 meteorological stations in Canada's Maritimes Provinces. Three proximity thresholds, δ_{25}, δ_{50}, and δ_{75}, are used, each defined by percentiles within the distribution of $d_{i,0}$. Combining the choice of the modulation function, the non-conformity score and the thresholds proximity, yields 12 scenarios of the proposed algorithm. These combinations likely result in different behaviors of the algorithm in defining the conformal band.

Performance of the proposed method is evaluated using:

Table 1. Algorithm Performances

	Cov_{α_L} %	Cov_{α_G} %	Width	InterScore	TimeT	TimeM
$A1_{\delta_{50},\mathcal{S}_{sup},\mathcal{D}_{sup}}$	88.58	8.57	3493.82	3988.14	19.74	0.56
$A1_{\delta_{25},\mathcal{S}_{sup},\mathcal{D}_{sup}}$	81.67	14.28	6605.34	9692.21	21.00	0.60
$A1_{\delta_{75},\mathcal{S}_{sup},\mathcal{D}_{sup}}$	72.86	2.85	2854.09	4861.55	14.03	0.40
$A1_{\delta_{50},\mathcal{S}_{sup},\mathcal{D}_{sqrt}}$	77.22	2.85	2993.69	4294.89	19.86	0.56
$A1_{\delta_{25},\mathcal{S}_{sup},\mathcal{D}_{sqrt}}$	68.03	11.42	5652.30	11016.10	20.99	0.59
$A1_{\delta_{75},\mathcal{S}_{sup},\mathcal{D}_{sqrt}}$	64.32	2.85	2465.61	5478.03	14.01	0.40
$A1_{\delta_{50},\mathcal{S}_{sqrt},\mathcal{D}_{sqrt}}$	79.11	14.28	8809.64	9705.91	19.77	0.56
$A1_{\delta_{25},\mathcal{S}_{sqrt},\mathcal{D}_{sqrt}}$	64.54	20.00	34739.03	37745.18	21.15	0.60
$A1_{\delta_{75},\mathcal{S}_{sqrt},\mathcal{D}_{sqrt}}$	68.96	17.14	6995.86	9463.04	13.96	0.39
$A1_{\delta_{50},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$	95.42	20.00	4319.03	4443.47	19.72	0.56
$A1_{\delta_{25},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$	85.12	17.14	10909.47	12088.76	21.03	0.60
$A1_{\delta_{75},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$	80.91	25.71	3455.53	4679.01	14.00	0.40

- coverage of $(1-\alpha)100\%$ prediction band $Cov_{\alpha_{Global}}$, i.e., the percentage of points of curve that are in the prediction band;
- the local coverage of the prediction band $Cov_{\alpha_{Local}}(t)$, i.e. the percentage of points of curves that are in the prediction band and it quantifies how many points are such that $Y_{v_i}(t) \in [I_l(t), I_u(t)]$;
- band width (Width), that give us an approximate margin of error (Width/2);
- the functional version of the interval score (IntScore), defined as:

$$S_\alpha(I(t), X_{s_i}(t)) = \int_T A(I(t), X_{s_i}(t))dt$$

with

$$A(I(t), X_{s_i}(t)) = (I_u(t) - I_l(t)) + \frac{2}{\alpha}(I_l(t) - X_{s_i}(t))_+ + \frac{2}{\alpha}(X_{s_i}(t) - I_u(t))_+$$

where $I(t) = [I_u(t), I_l(t)]$ is the predict band. In this analysis we use $\alpha = 0.5$.

Table 1 summarizes the primary findings. Let $Al_{\delta,\mathcal{S},\mathcal{D}}$ denote each scenario, where:

- $\delta \in \delta_{25}, \delta_{50}, \delta_{75}$ represents the threshold for partitioning the dataset into training and testing subsets.
- $\mathcal{S} \in \mathcal{S}sup, \mathcal{S}sqrt$ denotes one of the two modulation functions used.
- $\mathcal{D} \in \mathcal{D}sup, \mathcal{D}sqrt$ represents one of the non-conformity scores employed.

As we can see $A1_{\delta_{25},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$, $A1_{\delta_{50},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$ and $A1_{\delta_{75},\mathcal{S}_{sqrt},\mathcal{D}_{sup}}$ have the best performances according to the proposed performance indices Cov_{α_L} and Cov_{α_G}.

References

1. Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetric **21**, 224–239 (2010)
2. Diana, A., Romano, E., Irpino: Conformal prediction for spatio-functional regression models. In: Book of Short Papers SIS 2022. PEARSON (2022). ISBN 978-88-9192-736-1
3. Diana, A., Romano, E., Adzic, J.: Conformal prediction for functional kriging models. In: Book of Short papers 11th International Conference IES2023. IlViandante (2023). ISBN 979-12-803-3369-8
4. Diana, A., Romano, E., Irpino, A.: Distribution free prediction for geographically weighted functional regression models. Spat. Stat. **57** (2023). https://doi.org/10.1016/j.spasta.2023.100765
5. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal prediction bands for multivariate functional data. J. Multivariate Data Anal. (2022)
6. Franco-Villoria, M., Ignaccolo, R.: Bootstrap based uncertainty bands for prediction in functional kriging. Spat. Stat. **21A**, 130–148 (2017). https://doi.org/10.1016/j.spasta.2017.06.005
7. Giraldo, R., Delicado, P., Mateu, J.: Ordinary kriging for function-valued spatial data. Environ. Ecol. Stat. **18**, 411–426 (2011). https://doi.org/10.1007/s10651-010-0143-y
8. Goulard, M., Voltz, M.: Geostatistical interpolation of curves: a case study in soil science. In: Soares, A. (ed.) Geostatistics Troia 1992, pp. 805–816. Kluwer, Doedretcht (1993)
9. Ignaccolo, R., Mateu, J., Giraldo, R.: Kriging with external drift for functional data for air quality monitoring. Stoch. Environ. Res. Risk Assess. **28**, 1171–1186 (2014)
10. Mateu, J., Romano, E.: Advances in spatial functional statistics. Stoch. Env. Res. Risk Assess. **31**, 1–6 (2017)
11. Mateu, J., Giraldo, R. (eds.): Geostatistical Functional Data Analysis. Wiley, Hoboken (2021)
12. Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, New York (2005)

Causal Machine Learning for Medical Texts

Alessandro Albano(✉), Chiara Di Maria, Mariangela Sciandra, and Antonella Plaia

Department of Economics, Business, and Statistics, University of Palermo, Palermo, Italy
alessandro.albano@unipa.it

Abstract. Text analysis has become increasingly common in medical research, especially for tasks like patient diagnosis based on medical notes. However, most existing approaches do not account for causal relationships between words and diagnoses. This paper proposes a causal approach using the MIMIC-III dataset to identify words or word pairs that causally affect the probability of receiving a specific diagnosis. We employ causal forests to assess the impact of individual linguistic factors on patient outcomes while adjusting for potential confounders. Our analysis reveals significant causal relationships between specific terms in clinical notes and the presence of hypothyroidism diagnosis.

Keywords: Causal inference · text analysis · MIMIC-III · causal forest · hypothyroidism

1 Introduction

Recently, the use of texts as a source of information has rapidly grown in data analysis. Texts can be analysed in different ways, according to the research interest, like sentiment analysis, topic detection, and text exploratory analysis [9]. One of the fields where textual analysis is increasingly employed is medicine, primarily to determine diagnoses based on medical notes. The approaches most widely used to achieve this task are mainly associational, in the sense that predictions are based on the frequency of the co-occurrences of certain words in the documents.

In this work, we propose a causal approach to assess which words causally affect the probability of receiving a diagnosis. We analyse the MIMIC-III dataset, often considered a gold standard in the research community, for various tasks, such as predicting patient outcomes, identifying risk factors, and natural language processing. The main goal is to evaluate if the presence of a word (or pair of words) in the texts affects the probability that a diagnosis of interest belongs to the group of a patient's diagnoses. To do so, we used causal forests [1], which allow for the estimation of heterogeneous treatment effects.

The remainder of the paper is structured as follows: first, we describe the methods used to carry out the analysis; next, we move to the data description and analysis; and finally, the conclusions end the paper.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 34–39, 2025.
https://doi.org/10.1007/978-3-031-64346-0_6

2 Methods

Let us consider a setting where W and Y denote a treatment and a response variable of interest, respectively, while X is a set of observed variables which confound the relationship between W and Y. The average treatment effect (ATE) of W on Y is a unique measure quantifying the extent to which the treatment causally affects the response on average in the population. However, in many real-world settings, it is plausible that the causal effect of W on Y varies according to some subjects' characteristics, leading to the so-called effect heterogeneity. If there do not exist other confounders of W and Y except for those in X, the conditional average treatment effect (CATE), where the term "conditional" refers to the fact that the causal effects are estimated conditionally to other variables, is denoted by $\tau(X_i)$

$$Y_i = \tau(X_i)W_i + f(X_i) + \varepsilon_i, \tag{1}$$

where f is an unknown function describing the dependence of the outcome on the covariates, and ε is an error term having zero mean conditional on W and X. Let us denote by $e(x) = \mathbb{E}[W_i|X_i]$ and $m(x) = \mathbb{E}[Y_i|X_i]$ the propensity score and the conditional mean of Y, respectively. Equation (1) can be rewritten as

$$Y_i - m(x) = \tau(X_i)(W_i - e(x)) + \varepsilon_i. \tag{2}$$

The causal effects, $\tau(X_i)$, can then be obtained as a weighted regression of residuals on residuals, although it raises the problem of how to estimate $e(x)$ and $m(x)$, due to the fact that f is unknown. However, since the interest lies only in their predicted values, they can be obtained with any machine-learning method, such as random forests. We used the approach proposed by Athey et al. [1,2] to estimate causal effects using random forests, which combines the approaches proposed by [3] and [7], where the objective function to maximise is the difference of treatment effects in the subgroups. The CATE is subsequently obtained with a weighted regression as shown in Eq. (2), possibly using cross-validation, where the weights correspond to the adaptive neighbourhood weights (see [5]). As discussed in [1], a consistent estimator of the CATE is the ratio

$$CATE = \frac{\mathrm{Cov}(W, Y|X)}{\mathrm{Var}(W|X)}, \tag{3}$$

that coincides with the ATE in the case of a binary treatment, i.e. the difference $\mathbb{E}[Y|W = 1] - [Y|W = 0]$.

3 Data Analysis

The dataset used in this analysis is MIMIC-III (Medical Information Mart for Intensive Care III) [4], a freely accessible database comprising clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, from 2001 to 2012.

The dataset includes clinical notes such as discharge summaries, progress notes, and nursing notes. One of the key features of MIMIC III is its use of ICD-9 codes to categorise patient diagnoses. Each subject within the dataset is assigned one or more ICD-9[1] codes, with a maximum of 39 codes recorded for some subjects. It is worth noting that a subject is considered affected by a specific disease if at least one of their assigned ICD-9 codes corresponds to that specific condition.

In this paper, our focus lies on identifying key terms within discharge notes that exhibit a statistically significant causal relationship with selected diseases whose diagnosis is not straightforward. In particular, we focused on diagnoses of hypothyroidism. To perform causal inference, we employed the `grf` package [8] in `R`, which allows us to run causal forests as described in [1,2,10]. We estimated the treatment effect of individual words and word pairs from clinical notes, using weighted log odds [6] to select them. The occurrence of each word (or word-pair) was treated as a discrete treatment, while the set of other words except those used as treatment acted as potential confounders. This allowed us to isolate their impact on outcomes, adjusting for the other words in the text.

We started by analysing the effect of single words with high log-odds score magnitudes (either positive or negative). The words *thyroid* and *levothyroxine* presented, not surprisingly, large positive scores, and positive significant causal effects, as shown in Table 1. Indeed, the former refers to the organ mainly affected by hypothyroidism, while the latter to the active principle used to treat this disease. This means that if all the texts were "treated" so that they contain the word thyroid or levothyroxine, the probability that the patients would be diagnosed with hypothyroidism would increase.

Table 1. Conditional average treatment effects (CATE) and corresponding standard errors of the words 'thyroid' and 'levothyroxine'.

Word	CATE	Standard Error
thyroid	0.012	0.001
levothyroxine	0.263	0.003

We proceeded by analysing less trivial words, shown in Fig. 1. The word with the highest CATE is *tsh*, which is not surprising since it is the thyroid-stimulating hormone, a hormone produced by the pituitary gland that regulates the production of thyroid hormones from the thyroid gland. As regards biological sex, the observed effects for *female* and *woman* (0.010 and 0.012, respectively) signify that the diagnosis of hypothyroidism tends to be more likely among females in contrast to males, whose effect sizes for *male* and *man* are negative (-0.010 and -0.011). Notably, all these coefficients demonstrate statistical significance, being

[1] https://www.cdc.gov/nchs/icd/index.htm.

distinctly different from zero. This result is consistent with empirical evidence, since hypothyroidism is more common in women than men.

Several terms in the data frame are related to medical conditions. For instance, *afib* (atrial fibrillation), *hepatitis*, *chf* (congestive heart failure), and *uti* (urinary tract infection) represent specific health conditions. Except for hepatitis, all of them show a positive causal effect, and this may be because all these conditions are often associated with hypothyroidism, so they may concur with its diagnosis.

The term *baby* shows a negative but non-significant effect because of the low number of occurrences. The negative sign is, however, meaningful since hypothyroidism is a disease that typically affects adults. Finally, terms such as *head* and *cardiovascular* relate to anatomical structures or physiological systems, but their effects are not significant.

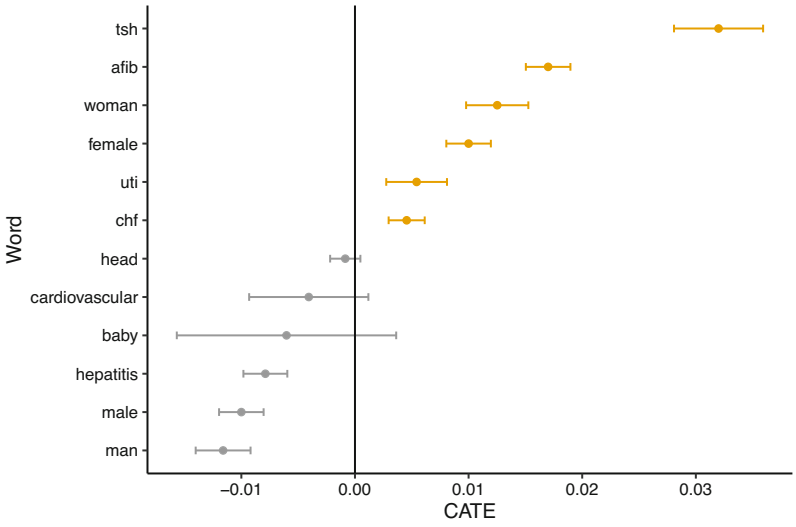


Fig. 1. Words' CATE estimates with 95% confidence interval.

Next, we investigated the effects of the co-occurrence of two words in the text. Drawing from our initial analysis, we focused on specific terms to explore changes in their causal effects when paired with words related to biological sex within the same text (refer to Fig. 2).

It is evident that the effects of word pairs related to males consistently exhibit lower CATEs compared to their female counterparts. In causal terms, this implies that if all texts were "treated" by adding word pairs related to biological sex, the probability of hypothyroidism diagnosis would increase more and more significantly for females than for males. Notably, the most noteworthy difference is observed for the word *tsh*, where the effect becomes positive also for males. A similar trend can be observed for *thyroid*, though to a lesser degree. Furthermore,

as regards *head*, the effect is positive for females and negative for males. This outcome is reasonable since the standalone term *head* did not previously exhibit a significant causal effect, thereby indicating the importance of sex-related words in determining the causal effect. Lastly, pairing *hepatitis* with males demonstrates an even more pronounced negative effect compared to the standalone use of hepatitis. Conversely, this association with females yields non-significant results.

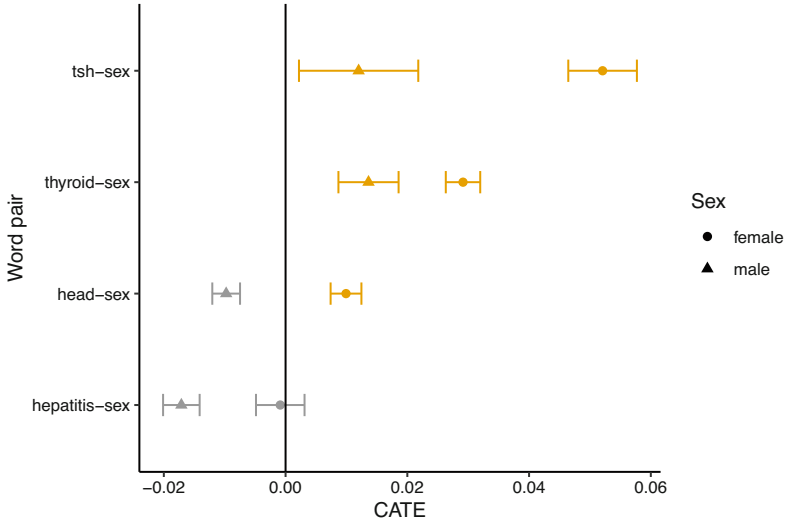


Fig. 2. Word pairs' estimates with 95% confidence interval.

4 Conclusions

This paper proposes a causal approach to analysing medical texts, specifically focusing on diagnosing diseases using the MIMIC-III dataset. The study identifies words or word pairs that causally affect the probability of receiving a specific diagnosis, employing causal forests to estimate heterogeneous treatment effects.

The results reveal a significant impact of biological sex on hypothyroidism diagnoses. Effect sizes for terms like *female* and *woman* are positive and significant, indicating a higher likelihood of hypothyroidism among females, while those for *male* and *man* are negative. Moreover, terms related to other medical conditions have large positive causal effects on hypothyroidism diagnoses. The analysis of word pairs yielded interesting results as well, showing for example that, pairing *tsh* and *thyroid* with biological sex, makes the causal effect positive. This is remarkable for the word *male*, which has a negative effect when it is considered singularly, but whose effects is positively influenced by the co-occurrence of words highly associated with hypothyroidism. These findings are statistically significant and align with empirical evidence.

Many research directions could be investigated in the future, for example, the application of this method to other diseases or the comparison of this causal approach with others proposed in the literature.

Acknowledgement. This research was partially funded by European Union - Next Generation EU. PRIN 2022 PNRR Project "A unified italian oral medicine and orthodontic language system: a prototype of Natural language processing application in healthcare" n. P202299ZNW CUP B53D23026050001.

References

1. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. Ann. Stat. **47**(2), 1148–1178 (2019)
2. Athey, S., Wager, S.: Estimating treatment effects with causal forests: an application. Obs. Stud. **5**(2), 37–51 (2019)
3. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
4. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**(1), 160035 (2016)
5. Lin, Y., Jeon, Y.: Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. **101**(474), 578–590 (2006)
6. Monroe, B.L., Colaresi, M.P., Quinn, K.M.: Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. Polit. Anal. **16**(4), 372–403 (2017)
7. Robinson, P.M.: Root-N-consistent semiparametric regression. Econometrica: J. Econ. Soc. 931–954 (1988)
8. Tibshirani, J., Athey, S., Sverdrup, E., Wager, S.: GRF: Generalized Random Forests. R package version 2.3.1 (2023)
9. Usai, A., Pironti, M., Mital, M., Mejri, C.A.: Knowledge discovery out of text data: a systematic review via text mining. J. Knowl. Manag. **22**(7), 1471–1488 (2018)
10. Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., Wager, S.: Evaluating treatment prioritization rules via rank-weighted average treatment effects. arXiv preprint arXiv:2111.07966 (2021)

A Positional Analysis of Ranking Procedures

José Luis García-Lapresta(✉) and Miguel Martínez-Panero

IMUVA, PRESAD Research Group, Departamento de Economía Aplicada, Universidad de Valladolid, Valladolid, Spain
lapresta@uva.es
https://eco.uva.es/lapresta/

Abstract. In this contribution, we analyze, from a formal perspective, the dense, standard, modified and fractional ranks. They assign a position to each alternative of a weak order from different points of view, generalizing the natural positions of linear orders. We show that the standard, modified and fractional ranks belong to a family of parameterized ranks. However, the dense rank has a different behavior, and it does not belong to that family. We provide some properties that all the mentioned ranks satisfy and we also show a characterization of the dense rank through two independent properties.

Keywords: dense rank · standard rank · modified rank · fractional rank

1 Introduction

In the setting of linear orders, it is straightforward to assign positions to the alternatives: that ranked first has position 1, the next one has position 2, and so on. However, in the setting of weak orders, where some alternatives may be indifferent, there are several possibilities to assign positions to the alternatives.

In this contribution, we have considered four position operators: the dense, standard, modified and fractional ranks, and also a family of parameterized ranks that includes the last three ranks as particular cases. The positions assigned by this family are convex combinations of the ones provided by the standard and modified ranks.

An interesting overview relating the dense rank to the mentioned ranks, with a suitable mathematical treatment, can be found in Vojnović [5, pp. 505–506].

In the frameworks of positional voting systems (see Gärdenfors [4] and García-Lapresta and Martínez-Panero [2]) and scoring rules (see Chevotarev and Shamis [1]), when these voting systems are extended from linear orders to weak orders, the way that positions are defined is crucial to generate a collective ranking of the alternatives.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 40–45, 2025.
https://doi.org/10.1007/978-3-031-64346-0_7

2 Dense, Standard, Modified and Fractional Ranks

In this section, we formally present the dense, standard, modified and fractional ranks.

2.1 Notation

Consider a finite set of alternatives $X = \{x_1, x_2, \ldots, x_n\}$, with $n \geqslant 2$. A *weak order* (or *complete preorder*) on X is a complete[1] and transitive[2] binary relation on X. A *linear order* on X is an antisymmetric[3] weak order on X. With $\mathcal{W}(X)$ and $\mathcal{L}(X)$, we denote the sets of weak and linear orders on X, respectively. Given $R \in \mathcal{W}(X)$, with P and I we denote the asymmetric and symmetric parts of R, respectively: $x_i \, P \, x_j$ if not $x_j \, R \, x_i$; and $x_i \, I \, x_j$ if ($x_i \, R \, x_j$ and $x_j \, R \, x_i$).

Given $R \in \mathcal{W}(X)$ and a permutation σ on $\{1, \ldots, n\}$, we denote by R^σ the weak order obtained from R by relabelling the alternatives according to σ, i.e., $x_i \, R \, x_j \;\Leftrightarrow\; x_{\sigma(i)} \, R^\sigma \, x_{\sigma(j)}$, for all $i, j \in \{1, \ldots, n\}$.

Given $R \in \mathcal{W}(X)$ and $Y \subseteq X$, the restriction of R to Y, $R|_Y$, is defined as $x_i \, R|_Y \, x_j$ if $x_i \, R \, x_j$, for all $x_i, x_j \in Y$. Note that $R|_Y \in \mathcal{W}(Y)$.

In turn, given a set Y, with $\#Y$ we denote its cardinality.

2.2 Positions in Weak Orders

We now introduce the notion of position operator. It assigns a position to each alternative in a weak order and allows us to add or withdraw alternatives along the process of assigning positions.

Definition 1. *Given a universe of alternatives U and $X \subseteq U$ finite, a* position operator O *assigns to each $R \in \mathcal{W}(X)$ a function $O_R : X \longrightarrow \mathbb{R}$. We say that $O_R(x_i)$ is the* position *of the alternative $x_i \in X$ in the weak order R.*

Once X has been fixed, given $R \in \mathcal{W}(X)$ and $x_i \in X$, we consider the number of alternatives dominated by x_i:

$$p_i = \#\{x_j \in X \mid x_i \, P \, x_j\}$$

and the number of alternatives indifferent to x_i:

$$q_i = \#\{x_j \in X \mid x_i \, I \, x_j\}.$$

Note that $p_i \in \{0, 1, \ldots, n-1\}$, $q_i \in \{1, 2, \ldots, n\}$ and $p_i + q_i \leqslant n$.

If $R \in \mathcal{L}(X)$, then $q_i = 1$ for every $i \in \{1, 2, \ldots, n\}$. However, if there are ties among alternatives, i.e., $R \in \big(\mathcal{W}(X) \setminus \mathcal{L}(X)\big)$, then $q_i > 1$ for some

[1] A binary relation R on X is complete if $x_i \, R \, x_j$ or $x_j \, R \, x_i$, for all $x_i, x_j \in X$.

[2] A binary relation R on X is transitive if ($x_i \, R \, x_j$ and $x_j \, R \, x_k$) implies $x_i \, R \, x_k$, for all $x_i, x_j, x_k \in X$.

[3] A binary relation R on X is antisymmetric if ($x_i \, R \, x_j$ and $x_j \, R \, x_i$) implies $x_i = x_j$, for all $x_i, x_j \in X$.

$i \in \{1, 2, \ldots, n\}$ and there will be $q_i - 1$ alternatives indifferent to x_i that are different to x_i.

Assigning positions to the alternatives in a linear order is a trivial task, as shown in the following definition.

Definition 2. *Given $R \in \mathcal{L}(X)$ and $x_i \in X$, the* natural rank *of x_i in R is defined as*

$$N_R(x_i) = \#\{x_j \in X \mid x_j \, R \, x_i\} = n - \#\{x_j \in X \mid x_i \, P \, x_j\} = n - p_i.$$

N_R assigns 1 to the alternative ranked first, 2 to the alternative ranked second, and so on.

Unlike the case of linear orders, it is not obvious how to assign positions to the alternatives in weak orders, where ties may appear.

2.3 Dense Rank

Surely, the dense rank is the most simple way to assign positions to the alternatives in a weak order. It assigns position 1 to all the alternatives belonging to the top indifference class, 2 to the second one, and so on.

Definition 3. *Given $R \in \mathcal{W}(X)$, for each $p \in \{0, 1, \ldots, n-1\}$, we consider the tier gathering all the alternatives that have p alternatives below,*

$$T_p = \{x_i \in X \mid p_i = p\},$$

and

$$T = \big\{p \in \{0, 1, \ldots, n-1\} \mid T_p \neq \emptyset\big\}.$$

Hereinafter, when we say tiers, we mean non-empty tiers i.e., T_p with $p \in T$.

Remark 1. If $R \in \big(\mathcal{W}(X) \setminus \mathcal{L}(X)\big)$, some T_p will be empty. However, always $T_0 \neq \emptyset$, hence $T \neq \emptyset$. What is more, for any $p \in T$, it holds that

$$\#T = \#\{p' \in T \mid p' > p\} + \#\{p' \in T \mid p' < p\} + 1,$$

where the second member corresponds to the number of tiers above and below T_p, plus 1 for T_p itself.

Definition 4. *Given $R \in \mathcal{W}(X)$ and $x_i \in X$, if $x_i \in T_p$, the* dense rank *of x_i in R is defined as*

$$D_R(x_i) = \#T - \#\{p' \in T \mid p' < p\} = \#\{p' \in T \mid p' > p\} + 1.$$

2.4 Standard, Modified and Fractional Ranks

Definition 5. *Given $R \in \mathcal{W}(X)$ and $x_i \in X$, the* standard rank *of x_i in R is defined as*

$$S_R(x_i) = n - p_i - (q_i - 1).$$

S_R assigns the best position in the indifference class, if the alternatives in that class were randomly linearized.

Definition 6. *Given $R \in \mathcal{W}(X)$ and $x_i \in X$, the* modified rank *of x_i in R is defined as*

$$M_R(x_i) = n - p_i.$$

M_R assigns the worst position in the indifference class, if the alternatives in that class were randomly linearized.

Definition 7. *Given $R \in \mathcal{W}(X)$ and $x_i \in X$, the* fractional rank *of x_i in R is defined as*

$$F_R(x_i) = n - p_i - \frac{1}{2} \cdot (q_i - 1).$$

F_R assigns the average position in the indifference class, if the alternatives in that class were randomly linearized. Since the natural positions are equispaced, the fractional rank is also the average of the standard and modified ranks:

$$F_R(x_i) = \frac{S_R(x_i) + M_R(x_i)}{2}.$$

Example 1. Consider the following weak order R on $X = \{x_1, x_2, \ldots, x_{12}\}$:

$$\begin{array}{c} x_3 \;\; x_8 \;\; x_{10} \\ x_5 \;\; x_{12} \\ x_{11} \\ x_1 \;\; x_4 \;\; x_7 \;\; x_9 \\ x_2 \;\; x_6 \end{array}$$

Table 1 includes the positions of the alternatives of X according to the dense, standard, modified and fractional ranks.

Table 1. Dense, standard, modified and fractional ranks

Alternatives	D_R	S_R	M_R	F_R
$x_3 \;\; x_8 \;\; x_{10}$	1	1	3	2
$x_5 \;\; x_{12}$	2	4	5	4.5
x_{11}	3	6	6	6
$x_1 \;\; x_4 \;\; x_7 \;\; x_9$	4	7	10	8.5
$x_2 \;\; x_6$	5	11	12	11.5

Standard, modified and fractional ranks are specific cases of the family of *parameterized ranks* defined as

$$\Lambda_R(x_i) = n - p_i - \lambda \cdot (q_i - 1),$$

with $\lambda \in [0, 1]$. Note that for $\lambda = 0, 0.5, 1$ we obtain the modified, fractional and standard ranks, respectively.

Remark 2. Consider the following weak order R on $X = \{x_1, x_2, x_3\}$:

$$\begin{array}{cc} x_1 & x_2 \\ & x_3 \end{array}$$

For every $\lambda \in [0, 1]$, we have $\Lambda_R(x_3) = 3$. However, $D_R(x_3) = 2$. Consequently, the dense rank does not belong to the family of parameterized ranks.

3 Properties

In this section we introduce some properties that position operators may satisfy. Excluding duplication, which is only satisfied by the dense rank, all other properties are fulfilled by the dense and parameterized ranks.

Definition 8. *Let O be a position operator and $O_R : X \longrightarrow \mathbb{R}$ the function that assigns a position to each alternative of X in the weak order $R \in \mathcal{W}(X)$. We say that the position operator O satisfies the following conditions, when they are fulfilled for all $X \subseteq U$ and $R \in \mathcal{W}(X)$:*

1. Equality*: $x_i \, I \, x_j \;\Rightarrow\; O_R(x_i) = O_R(x_j)$, for all $x_i, x_j \in X$.*
2. Monotonicity*: $x_i \, R \, x_j \;\Leftrightarrow\; O_R(x_i) \leqslant O_R(x_j)$, for all $x_i, x_j \in X$.*
3. Neutrality*: $O_{R^\sigma}(x_{\sigma(i)}) = O_R(x_i)$ for every permutation σ on $\{1, \ldots, n\}$.*
4. Sequentiality*: If $R \in \mathcal{L}(X)$, then $O_R(x_i) = N_R(x_i)$, for every $x_i \in X$.*
5. Truncation*: $O_{R|_{X \setminus T_0}}(x_i) = O_R(x_i)$, for every $x_i \in X \setminus T_0$.*
6. Duplication*: if, whenever $R \in \mathcal{W}(X)$, $R' \in \mathcal{W}(X')$, with $X' = X \cup \{x_{n+1}\}$ such that $x_{n+1} \notin X$, $R'|_X = R$ and $x_{n+1} \, I' \, x_j$ for some $x_j \in X$, then $O_{R'}(x_i) = O_R(x_i)$ for every $x_i \in X$ and $O_{R'}(x_{n+1}) = O_{R'}(x_j)$.*

Proposition 1

1. *If a position operator satisfies duplication, then it also satisfies neutrality.*
2. *If a position operator satisfies neutrality, then it also satisfies equality.*
3. *If a position operator satisfies monotonicity, then it also satisfies equality.*
4. *The dense and all the parameterized ranks satisfy neutrality (hence, equality), monotonicity, truncation and sequentiality.*
5. *The dense rank satisfies duplication, but the parameterized ranks do not.*

In the following result we show a characterization of the dense rank.

Theorem 1 *(García-Lapresta and Martínez-Panero [3]). A position operator O satisfies sequentiality and duplication if and only if, for each $X \subseteq U$ finite and $R \in \mathcal{W}(X)$, the function $O_R : X \longrightarrow \mathbb{R}$ assigns to each $x_i \in X$ the dense rank of x_i in R.*

Remark 3. The result appearing in Theorem 1 does not use the compelling requirement of monotonicity. Nevertheless, other axiomatizations of the dense rank would be possible by incorporating this property. Another characterization of the dense rank can be obtained by keeping sequentiality, adding monotonicity and relaxing duplication (see García-Lapresta and Martínez-Panero [3]).

4 Concluding Remarks

Proposition 1 establishes some properties that the dense and parameterized rank satisfy. In Theorem 1 we have shown one of the characterizations of the dense rank included in García-Lapresta and Martínez-Panero [3]. As further research, we shall try to characterize the family of parameterized ranks and, specifically, the standard, modified and fractional ranks.

Acknowledgments. The financial support of the Spanish *Ministerio de Ciencia e Innovación* (project PID2021-122506NB-I00) is acknowledged.

References

1. Chevotarev, P.Y., Shamis, E.: Characterizations of scoring methods for preference aggregation. Ann. Oper. Res. **80**, 299–332 (1998)
2. García-Lapresta, J.L., Martínez-Panero, M.: Positional voting rules generated by aggregation functions and the role of duplication. Int. J. Intell. Syst. **32**, 926–946 (2017)
3. García-Lapresta, J.L., Martínez-Panero, M.: Two characterizations of the dense rank. J. Math. Econ. 102963 (2024)
4. Gärdenfors, P.: Positionalist voting functions. Theor. Decis. **4**, 1–24 (1973)
5. Vojnović, M.: Contest Theory: Incentive Mechanisms and Ranking Methods. Cambridge University Press, Cambridge (2016)

Non-metric Unfolding via Copula

Marta Nai Ruscone[1(✉)] and Antonio D'Ambrosio[2]

[1] University of Genoa, Via Vivaldi, 5, 16126 Genoa, Italy
marta.nairuscone@unige.it
[2] University of Naples Federico II, Via Cinthia, M.te S. Angelo, 80125 Napoli, Italy

Abstract. An effective procedure to avoid degeneracies in multidimensional unfolding for preference rank data is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using copula-based association measures among rankings (individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). Our proposal is able to both recover the order of the preferences and reproduce the position of both rankings and objects in a geometrical space. Application on real datasets show that our procedure returns non-degenerate unfolding solutions.

Keywords: Copula · Multidimensional scaling · Unfolding

1 Copula

Copula are functions that join multivariate distribution functions to their marginal distribution functions [8]. They describe the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate normal distribution. Each copula is related to the most important measures of dependency: the Pearson correlation coefficient and the Spearman ρ correlation coefficient. The Spearman's ρ coefficient (see [8] pp. 169–170 for the definition of the ρ correlation coefficient for continuous random variables) measures the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ, then the Spearman ρ correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3. \tag{1}$$

For continuous random variables it is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [5].

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 46–49, 2025.
https://doi.org/10.1007/978-3-031-64346-0_8

2 Unfolding as a Special Case of Multidimensional Scaling on Copula-Based Association Between Rankings

Unfolding, originally formulated by Coombs [3] for the analysis of the two-mode preference choice data, is a technique that allows the estimation of two configurations usually representing the coordinates for a set of m individuals and a set of n objects on the basis of proximity values between them, typically expressing preferences of each individual over each object.

Therefore unfolding applies multidimensional scaling [4] to an off-diagonal $n \times m$ matrix, usually representing the scores (or the rank) assigned to a set of m items by n individuals or judges [1]. Using of either scores or rankings traditionally discriminates between metric and non-metric unfolding.

The goal is to obtain two configuration of points representing the position of the judges (X) and the items (Y) in a reduced geometrical space. Each point representing the individuals is considered as an ideal point so that its distances to the object points correspond to the preference scores [3].

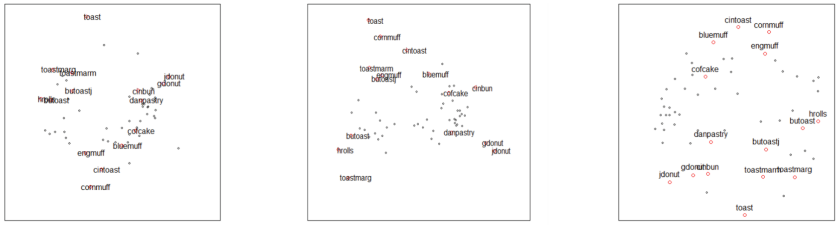
Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions, i.e., solutions that return excellent badness of fit measures but not graphically interpretable at all.

To tackle the problem of degenerate solutions, several proposals have been presented in the literature [1]. By following the approach introduced by [9], we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms.

Let Γ be the original $m \times n$ original preference data matrix. In order to augment the data matrix we add to this n additional rows, one for each of the n objects, that correspond to tied rankings representing the jth item, $j = 1, \ldots, n$. As a result, a new $N \times n$ Γ^* matrix is obtained, with $N = n + m$. Then we use copula-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings), obtaining in fact a $N \times N$ dissimilarity matrix to be analyzed with any MDS algorithm.

3 An Application on a Real Data Set

Figure 1 shows a comparison between the Unfolding solutions of PRESCAL [2], which actually is the most used algorithm for Unfolding analysis, and our proposal by using the Spearman ρ correlation coefficient via copula on the breakfast data set. Green and Rao [6] collected 42 rankings of 15 objects by asking 21 students and their wives to order 15 breakfast items in terms of their preference.



(a) Row-conditional Prefscal solution (b) Unconditional Prefscal solution (c) Spearman augmented unfolding

Fig. 1. Unfolding solutions for breakfast data. Breakfast items are labeled as follow: Toast pop-up; Buttered toast; English muffin and margarine; Jelly donut; Cinnamon toast; Blueberry muffin and margarine; Hard rolls and butter; Toast and marmalade; Buttered toast and jelly; Toast and margarine; Cinnamon bun; Danish pastry; Glazed donut; Coffee cake; Corn muffin and butter.

PREFSCAL works by setting two penalties on a modified loss function in such a way to guarantee non degenerate solutions. A possible drawback of this algorithm is that it is not always clear how set the penalty terms. In fact the user must make attempts in order to find the right solution.

The figure emphasizes that the solution of our procedure is not degenerate and it is comparable with the one of PREFSCAL, especially with its unconditional output. It is normal that our output looks like the unconditional PREFSCAL solution because we propose a solution that, depending on how we defined the dissimilarity matrix, is unconditional as well.

4 Concluding Remarks

In this work we propose an unfolding algorithm based on the augmentation of the data matrix and a copula-based association between rankings. The shown result highlights that our proposal produces non-degenerate unfolding solutions that are comparable with the ones obtained with PREFSCAL. With respect to PREFSCAL, any parameter must be a priori chosen by the user. On the other hand PREFSCAL always guarantees non degenerate solutions. Unfolding, originally formulated by Coombs [3] for the analysis of the two-mode preference choice data, is a technique that allows the estimation of two configurations usually representing the coordinates for a set of m individuals and a set of n objects on the basis of proximity values between them, typically expressing preferences of each individual over each object.

References

1. Borg, I., Groenen, P.: Modern Multidimensional Scaling. Theory and Applications. Springer, New York (1997)
2. Busing, F.M.T.A., Groenen, P.J.K., Heiser, W.J.: Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika **70**(1), 71–98 (2005)
3. Coombs, C.H.: Psychological scaling without a unit measurement. Psychol. Rev. **57**, 145–158 (1998)
4. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling. Chapman & Hall, London (1994)
5. Embrechts, P., McNeil, A.J., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. Risk Management: Value at Risk and Beyond, pp. 176–223 (1998)
6. Green, P.E., Rao, V.R.: Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. Holt. Rinheart and Winston, New York (1972)
7. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York (1990)
8. Nelsen, R.B.: An Introduction to Copulas. Springer Series in Statistics, Springer, New York (2013)
9. Van Deun, K., Heiser, W., Delbeke, L.: Multidimensional unfolding by nonmetric multidimensional scaling of spearman distances in the extended permutation polytope. Multivar. Behav. Res. (2007). https://doi.org/10.1080/00273170701341167

Analysing Differences in Ranking Distributions

Rosaria Simone(✉)

University of Naples Federico II, Naples, Italy
rosaria.simone@unina.it

Abstract. The contribution presents a pilot exercise on the analysis of differences in ranking data combining non-parametric and parametric methods. Suitable hypothesis testing procedures for discrete data are considered to identify significance differences in response profiles determined through benchmark classification trees. For marginal rankings, these response profiles can be further parameterized with a mixture of discretized Beta distribution to model polarization towards the extremes positions and floatation in between. As a by-product of the exercise, a new classification tree for qualitative outcomes is advanced that allows to disclose both global and local differences in response distributions.

Keywords: Ranking data · Classification trees · Hypothesis testing · Finite Mixture Models

1 Motivation and Illustrative Data Example

Rankings are frequently used to survey preference data when the interest lies in understanding relative preferences of different survey items. They are multivariate data, which can be analysed with two different univariate approaches:

- dealing with marginal ranking distributions, thus considering the ranks assigned to each item as a rating on a preference scale; an ordinal variable corresponds to each item.
- dealing with ranking positional variables on the nominal scale, thus identifying - for each respondent - which item has been ranked first, which one second, and so on, up to the least preferred item. A nominal variable then corresponds to each position in the ranking.

The contribution aims at proposing a combination of parametric and non-parametric methods to analyse ranking data and disclose relevant significant differences in the response profiles obtained with classification trees.

For marginal rankings, a mixture of discretized Beta distributions, introduced in [3], is considered to parameterize polarization towards the extreme positions and floatation in between. In order to derive response profiles in terms of subjects characteristics, the corresponding model-based tree however would be rather

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 50–56, 2025.
https://doi.org/10.1007/978-3-031-64346-0_9

cumbersome and not parsimonious, for both implementation and interpretation, since the most general specification foresees covariates linked to each of the parameters. For this reason, we present a strategy to obtain response profiles for marginal rankings, characterized in terms of mixture parameters, by first running a conditional inference tree [1]. This benchmark is chosen since it yields unbiased trees relying on permutation tests for the partitioning process, thus guaranteeing that the (conditional) children distributions are significantly different from that of the parent node. However, this method does not provide information about local differences in the distributions of the descendant nodes: the same issue can be raised for other non-parametric classification trees.

We deal with this argument by considering testing procedures for discrete data implemented in the `R` package WRS2 [2] to check whether significant differences occur at any response category for the terminal nodes of a given classification tree. For the significant response profiles so obtained for marginal rankings, the chosen model can be then fitted to parameterize them in a comparative perspective. In the end, a new proposal for a classification tree for general qualitative outcomes is advanced, called *significance tree*, that can be used also to disclose different response profiles for positional ranking variables.

2 Combining Parametric and Non-Parametric Methods for Marginal Rankings

Assume that R is the variable collecting preference data over a support with m categories: R can be either a marginal ranking, or a categorical outcome reporting the item set in a given position in a ranking.

Section 2.1 briefly lists the hypothesis tests chosen to investigate global and local significant differences in the response distributions induced by a classification tree[1], whereas Sect. 2.2 introduces the statistical model that will be fitted locally to parameterize the corresponding response profile.

2.1 Hypothesis Testing for Differences in Discrete Distributions

In order to test for differences in discrete distributions, we consider some procedures available within the `R` package `WRS2` [2]:

- Discrete Anova hinges on Chi-squared tests to assess if the conditional distributions $R|D = 0$ and $R|D = 1$ are equal, given a binary split D;
- The `binband` test resorts to the Storer-Kim method to compare binomial proportions [4], correcting for multiple testing.

[1] The following pre-pruning conditions were set to grow all the classification trees: minimum number of observations in each descendant of a split (`minbucket`=50); minimum number of observations in a node to attempt a split (`minsplit`=100); maximum depth of the tree (`maxdepth`=3).

2.2 OFS Mixture of Discretized Beta Distributions

The class of OFS[2] finite mixtures of discretized Beta distribution *db* to model ordered data has been proposed in [3]. The most general model is specified as:

$$Pr(R = r|\theta) = \delta_1\, db_r(\alpha_1, 1) \;+\; \delta_2 db_r(\alpha_2, \beta_2) + \delta_3\, db_r(1, \beta_3), \quad r = 1, \ldots, m,$$

where:

- the *opposition* component $db(\alpha_1, 1)$ models polarization towards the first category of the support, with $\alpha_1 \in (0, 1)$ and mode at $c = 1$. The lower α is, the stronger is the polarization towards the mode;
- the *support* component $db(1, \beta_3)$ models polarization towards the last category, with $\beta_3 \in (0, 1)$, and mode at $c = m$: the lower β_3 is, the stronger is the polarization towards the mode;
- the *floatation* component $db(\alpha_2, \beta_2)$ has mode at $c \neq 1, m$, with $\alpha_2, \beta_2 > 1$. Its skewness and kurtosis indicate direction and intensity of the floatation.

3 Profiling Sport Preferences via Marginal Ranking Distributions

The pilot exercise will exploit a survey investigating preferences in sports as illustrative data example[3]. Rankings (without ties) for the 8 most popular sports were collected for a total of $n = 647$ observations.

For the sake of illustration we focus on marginal rankings obtained by soccer. Figure 1 displays the structure of the `ctree` (left panel) and the local OFS fit (right panel): see Table 1 for corresponding parameter estimates, allowing comparisons of nodes in terms of polarization towards the first and the last positions (in terms of both size and strength) and floatation among them.

Table 1. OFS model fitted to terminal nodes of `ctree` on marginal rankings for soccer.

Node	δ_1	δ_2	α_1	α_2	β_2	β_3
Root	0.401	0.475	0.178	0.795	1.655	0.325
Node 2	0.569	0.295	0.070	0.624	1.955	0.272
Node 4	0.306	0.608	0.470	2.401	2.712	0.312
Node 5	0.901	0.099	0.637	3.382	2.483	0.003

Table 2 reports results of the `binband` test for comparing the frequencies of each response category for pairs of nodes of the `ctree`. It is found that no

[2] The acronym OFS stands for opposition-floatation-support.

[3] The survey was planned and run in 2016 by students enrolled to the master degree program in Statistics at the University of Naples Federico II.

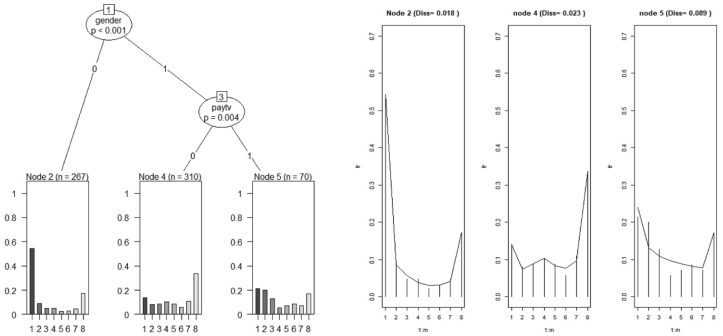


Fig. 1. `ctree` for marginal rankings of soccer (left); local OFS fit (right).

significant local differences - after correcting for multiple testing - are found when comparing nodes 4 and 5: thus, this could be pruned, or different stopping rules should be set. This circumstance motivates the introduction of a parsimonious partitioning algorithm that, regardless of the tuning of stopping rules, produces a classification tree in which each split is justified by the existence of significant differences in frequencies in at least one category, correcting for multiple testing.

Table 2. p-values and critical values for multiple testing of the `binband` procedure to compare `ctree` nodes for soccer marginal rankings: significant diffrences are in bold.

Category	Nodes 4, 5		Nodes 2,4		Nodes 2, 5		Nodes 2, 3	
	p.value	p.crit	p.value	p.crit	p.value	p.crit	p.value	p.crit
1	0.180	0.008	**0.001**	0.008	**0.001**	0.006	**0.001**	0.007
2	0.027	0.007	0.700	0.050	0.041	0.007	0.620	0.050
3	0.390	0.010	0.076	0.017	0.098	0.008	0.036	0.017
4	0.420	0.012	0.017	0.012	0.790	0.025	0.036	0.012
5	0.830	0.050	**0.001**	0.007	0.250	0.012	**0.002**	0.008
6	0.520	0.025	0.120	0.025	0.210	0.010	0.074	0.025
7	0.520	0.017	**0.005**	0.010	0.520	0.017	**0.010**	0.010
8	0.010	0.006	**0.001**	0.006	0.970	0.050	**0.001**	0.006

4 The Significance Tree for Qualitative Responses

The *significance tree* relies on the following recursive partitioning algorithm that embeds global and local hypothesis testing for discrete distributions:

– Step 1: Run discrete Anova for each candidate splitting variable D to test the global hypothesis that the conditional distributions $R|D = 0$ and $R|D = 1$ are identical, assuming an underlying multinomial distribution for both;

– Step 2: Among the candidate splits for which discrete Anova is significant at a given level, select the one for which the `binband` test for comparisons of (multiple) binomial proportions (one for each response category) for the two descendants is the most significant. With more details, the procedure identifies the binary split D^* with the maximum number of significant differences (in case there are more than one of such cases, the one with the strongest significant difference is selected).
– Iterate the procedure until a stopping rule is met[4].

4.1 The Significance Tree for Marginal Ranking Distributions

Figure 2 displays the structure of the significance tree (left) and the distributions at the terminal nodes, with highlights on categories for which the procedure has locally found significant differences; the estimated OFS model is superimposed to the observed response distribution. In this regard, Table 3 reports parameter estimates at each terminal node. Compared to Fig. 1, the significance tree seems to entail better local fitting performance for OFS model and more discriminating response profiles.

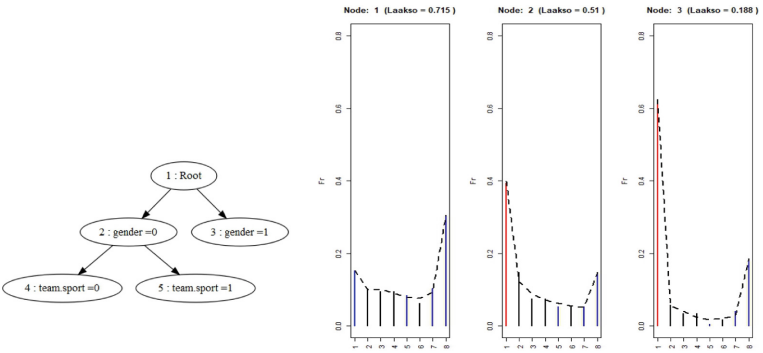


Fig. 2. The significance tree to profile marginal rankings of soccer

Table 3. Parameter estimates of OFS model fitted to terminal nodes of the significance tree for soccer marginal rankings ($\mathcal{L} = 0.593$ for the root node).

Node	δ_1	δ_2	α_1	α_2	β_2	β_3
3	0.227	0.615	0.341	0.977	1.657	0.352
4	0.888	0.112	0.385	3.179	3.401	0.033
5	0.595	0.292	0.001	0.980	2.186	0.217

[4] For comparative purposes, the same pre-pruning conditions used for the other classification trees considered in the analysis were set.

4.2 The Significance Tree for Positional Ranking Variables

The significance tree can be applied also to obtain response profiles for ranking positional variables. For instance, Fig. 3 displays the significance tree and the node distributions for the most preferred sport, whose frequency distribution is reported in Table 4:

Table 4. Frequency distribution of the positioning variable 'First' ($\mathcal{L} = 0.664$).

	Basketball	Boxing	Soccer	Cycling	Jogging	Swimming	Volleyball	Tennis
First	44	57	203	35	100	95	80	33

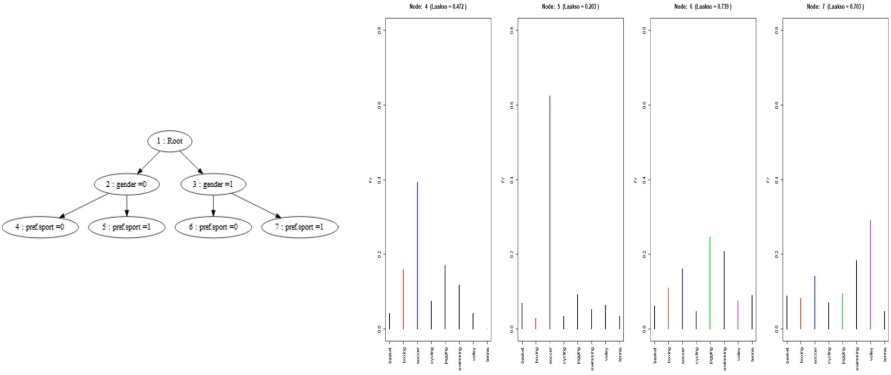


Fig. 3. Significant tree for the most preferred sport (left); nodes distributions (right)

5 Further Developments

The ultimate goal of the analysis is the proposal of a general partitioning method for qualitative outcomes that allows to identify both global and local relevant differences in response profiles on the basis of suitable statistical tests and of flexible statistical models for discrete data.

With respect to the state of the art, a further competitor would be the benchmark CART algorithm: even if such procedure does not allow to disclose possible significant differences that may exist locally, it should be considered together with `ctree` as competitor in assessing the predictive performance of the proposed significance tree for general qualitative outcomes.

References

1. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**(3), 651–674 (2006)
2. Mair, P., Wilcox, R.: Robust statistical methods in R using th WRS2 package. Behav. Res. Methods **52**, 464–488 (2020)
3. Simone, R.: On finite mixtures of discretized beta model for ordered responses. TEST **31**, 828–855 (2022)
4. Storer, B.E., Kim C.: Exact properties of some exact test statistics for comparing two binomial proportions. J. Am. Stat. Assoc. **85**(409), 146–155 (1990)

Mapping Well-Being Through a Mixture-of-Experts Fay-Herriot Model

Aldo Gardini[1], Silvia De Nicoló[1], and Enrico Fabrizi[2(✉)]

[1] Department of Statistical Sciences, Universitá di Bologna, Bologna, Italy
{aldo.gardini,silvia.denicolo}@unibo.it
[2] DISES & DSS, Universitá Cattolica del S. Cuore, Milan, Italy
enrico.fabrizi@unicatt.it

Abstract. Our research is motivated by estimation of the per capita wealth index in Bangladeshi upazilas, integrating data from the Demographic and Health Survey along with remote sensing covariates, in a small area estimation framework. The popular Fay-Herriot model shows relevant limitations when applied to our data, as it fails to manage adequately two (or more) distinct regimes in the data generating process, such as those implied by the rural/urban divide that characterizes many low and middle income countries. We extend the Fay-Herriot model through a Mixture-of-Experts, where areas are classified into groups within the estimation process. This adds flexibility while keeping relevant properties of the predictors such as design consistency and easy interpretation. In addition, this family of models defines the mixing probabilities through a logistic regression, turning out to be particularly convenient in the applied setting.

Keywords: Bayesian inference · DHS survey · small area estimation · well-being indicators

1 Introduction

Estimates of socio-economic indicators are often needed for small subsets of a population in order to monitor their geographical and/or social distribution. The primary data source for estimation is often a sample survey, which typically fall short of providing reliable estimates for small sub-populations (small area) because of limited area-specific sample sizes. To improve the reliability of these estimates, small area estimation methods integrate survey data with other data sources. The different data sources can be integrated at the unit/household level or at the area of interest level as we do in this research (see [5], Ch. 6 and 7).

The Fay-Herriot (FH) [2] model is the simplest and most popular area-level model. One distinguishing feature of this model is the linear regression that links the area-level target parameter to auxiliary information. This entails the assumption of a single intercept and regression slopes for the whole ensemble of

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 57–62, 2025.
https://doi.org/10.1007/978-3-031-64346-0_10

areas being considered. Such an assumption can be very restrictive when studying problems characterised by a relationship between response and covariates that varies across the areas.

The need for an extension of FH model is motivated by an application. Specifically, we consider the estimation of the per capita wealth index (WI) in Bangladeshi upazilas, i.e., small administrative subdivisions. Direct survey estimates are obtained from the Demographic and Health Survey (DHS), while auxiliary information comes mostly from remote sensing (RS) satellite images. A rural/urban divide in the WI distribution is typically observed for developing countries, induced by markedly different economies. From the statistical perspective, this can translate into the existence of separate data generating processes. This step could be nontrivial in area-level models as the rural/urban classification of territories does not perfectly overlap with administrative boundaries. As a consequence, the areas cannot be classified a priori.

In this paper, we face this challenge by proposing a Bayesian Mixture-of-Experts (MoE) model [3] that extends the FH model by allowing for the existence of different groups of areas each characterised by distinct predictors. An important feature of MoE is the modelling of the area-specific probability of belonging to a given group (known also as *gating* probabilities) through a logistic regression. Each model component (i.e. the *expert*) is a distinct Fay-Herriot model, namely allowing for different coefficients and/or area-specific random effects.

Although the framework we are dealing with relates to previous proposals in the small area literature (e.g., [6]), it presents distinct novelties, especially in the way we model the membership probabilities, which are allowed to vary across areas. Moreover, it benefits the estimation process in various ways. In the first place, the MoE approach enhance model flexibility while keeping important properties associated with the FH model, such as design consistency and easy interpretation of the predictors. Secondly, our proposal incorporates both the clustering and the small area estimation steps into a unique procedure. In this way, it inherently accounts for the uncertainty involved in the clustering step in the resulting estimator. Eventually, the MoE provides a distinct advantage when prediction has to be carried out for out-of-sample (OOS) areas, as disposing of non-constant group probabilities can be useful in this step.

In this short paper, we present the methodology, overlooking formal proofs, and provide a sketch of the application results. A thorough comparison of various methods using simulation is omitted for brevity.

2 Methodology

Let $\mathcal{U}$ indicates a finite population constituted by N individuals that is partitioned into D small areas $\mathcal{U}_d$. Each sub-population $\mathcal{U}_d$ includes N_d units, such that $N = \sum_{d=1}^{D} N_d$. A sample of overall size n is drawn from $\mathcal{U}$ through a complex survey scheme; area-specific sample sizes are defined n_d so that $n = \sum_{d=1}^{D_{IS}} n_d$. We denote the WI we are interested in as Y, so that our target of inference is

defined as $\theta_d = N_d^{-1} \sum_{j=1}^{N_d} y_{dj}$; where y_{dj} denotes the observed WI of individual j belonging to area d.

The Hájek type direct estimator we consider, whose definition includes published survey weights w_{dj} is given by

$$\hat{\overline{Y}}_d = \frac{\sum_{j=1}^{n_d} w_{dj} y_{dj}}{\sum_{j=1}^{n_{dj}} w_{dj}}, \quad d = 1, \ldots, D_{IS}. \tag{1}$$

In line with the ordinary FH model we specify the following sampling model:

$$\hat{\overline{Y}}_d | \theta_d, S_d^2 \overset{ind}{\sim} \mathcal{N}\left(\theta_d, S_d^2\right), \ d = 1, \ldots, D_{IS}. \tag{2}$$

where D_{IS} is the number of in-sample areas for which $n_d > 0$. Moreover, S_d is the standard error associated to the direct estimate $\hat{\overline{Y}}_d$ and will be treated as known. Actually, we computed it the product of two terms: the standard error under the assumption of simple random sample and a design effect that accounts for all design features (and namely the variable weights and the clustering of observations).

The linking level of the standard FH model that assumes normality is given by:

$$\theta_d | \beta_0, \boldsymbol{\beta}, \sigma_u \overset{ind}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_d^T \boldsymbol{\beta}, \sigma_u^2), \ \forall d; \tag{3}$$

where $\mathbf{x}_d \in \mathbb{R}^P$ contains the values of P auxiliary variables registered for area d. To complete the Bayesian specification, prior distributions must be specified for both the model parameters and hyperparameters. A fully proper prior setting is adopted. We specify a standard half-normal distribution for the scale parameter σ_u. Following the guidance of [4], we set the prior for the intercept as $\beta_0 \sim \mathcal{N}(m_y, 2.5^2 s_y^2)$, where m_y and s_y^2 represent the mean and variance of the direct estimates, respectively. Lastly, for the regression coefficients, we impose $\beta_p \sim \mathcal{N}(0, 2.5^2), \forall p$, provided that all the covariates are standardised.

We can express the linking level of our FH model extended to include a mixture of experts (FH-MoE) as:

$$\begin{aligned} \theta_d | \beta_{01}, \beta_{02}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_{u1}, \sigma_{u2}, z_d \overset{ind}{\sim} & \ \mathbf{1}(z_d = 1) \mathcal{N}\left(\beta_{01} + \mathbf{x}_d^T \boldsymbol{\beta}_1, \sigma_{u1}^2\right) \\ & + \ \mathbf{1}(z_d = 2) \mathcal{N}\left(\beta_{02} + \mathbf{x}_d^T \boldsymbol{\beta}_2, \sigma_{u2}^2\right), \ \forall d; \end{aligned} \tag{4}$$

where z_d is a dichotomous latent variable that determines the cluster label for area d, assuming values 1 or 2. In addition, $\mathbf{1}(z_d = k)$ is an indicator variable that assumes value 1 if $z_d = k$ and zero otherwise, and each model component k is characterised by distinct coefficients $[\beta_{0k}, \boldsymbol{\beta}_k^T]^T$ and scale σ_{uk}. We apply the same prior distribution used for the FH model to the component-specific parameters. Likewise, we assume a half-normal prior for the random effects scale σ_v. Additionally, independent Gaussian distributions with a mean of zero and a standard deviation of 2.5 are set for the coefficients.

The probability that z_d assumes value 1 is denoted as $\mathbb{P}[z_d = 1 | \mathbf{x}_d^*] = \pi(\mathbf{x}_d^*)$, where $\mathbf{x}_d^* \in \mathbb{R}^{\tilde{P}}$ is a vector of covariates possibly different to the set used in

modelling the location $\mathbf{x}_d$. In the MoE framework, this quantity is also defined as the gating network and can be interpreted as the prior probability that area d is assigned to the first mixture component. Specifically, for $\pi(\mathbf{x}_d^*)$, we assume a logistic regression:

$$\log\left(\frac{\pi(\mathbf{x}_d^*)}{1-\pi(\mathbf{x}_d^*)}\right)\Bigg|\gamma_0,\boldsymbol{\gamma},v_d=\gamma_0+\mathbf{x}_d^{*T}\boldsymbol{\gamma}+v_d,\ \forall d; \tag{5}$$

where γ_0 is the intercept, $\boldsymbol{\gamma}\in\mathbb{R}^{\tilde{P}}$ is the vector of regression coefficients and $v_d|\sigma_v \overset{ind}{\sim} \mathcal{N}(0,\sigma_v^2),\ \forall d$, is an area-specific random effect.

Once we set $\boldsymbol{\Psi}=(\boldsymbol{\Psi}_1^T,\boldsymbol{\Psi}_2^T)$, where $\boldsymbol{\Psi}_k=(\beta_{0k},\boldsymbol{\beta}_k,\sigma_{uk})$, for $k=1,2$, the predictor minimizing a squared loss can be expressed as

$$\begin{aligned}\mathbb{E}\left[\theta_d|\boldsymbol{\Psi},\pi\left(\mathbf{x}_d^*\right),\hat{\overline{Y}}_d\right]&=(1-\tilde{\eta}_d)\hat{\overline{Y}}_d+\pi_P\left(\mathbf{x}_d^*\right)\eta_{d1}(\beta_{01}+\mathbf{x}_d^T\boldsymbol{\beta}_1)\\&\quad+(1-\pi_P\left(\mathbf{x}_d^*\right))(1-\eta_{d2})(\beta_{02}+\mathbf{x}_d^T\boldsymbol{\beta}_2).\end{aligned} \tag{6}$$

where $\tilde{\eta}_d$ is the overall shrinkage factor we can write as

$$\tilde{\eta}_d=\pi_P(\mathbf{x}_d^*)\eta_{d1}+\left[1-\pi_P(\mathbf{x}_d^*)\right]\eta_{d2} \tag{7}$$

and $\eta_{dk}=S_d^2/(\sigma_{uk}^2+S_d^2)$, $k=1,2$ is the expert-specific shrinkage factor.

We remark that we illustrated a mixture with two components, due to features of the motivating application and the tendency for small area models to be parsimonious. However, it is worth mentioning that the results presented here can be readily extended to a model with multiple components if needed.

3 Application

We used DHS survey data for Bangladesh, 2014 wave. A thorough description of these data and the remote sensing (RS) data used as covariates, can be found in [1] and is omitted here for brevity. We only mention that the considered RS covariates include demographics (population density and its Geary index, child to woman and male to female ratios) along with development indicators (night time light radiance, proportion of built and agricultural areas, land use, time health, and other facilities).

These many covariates suffer from quasi-multicollinearity, a problem that we solve using principal components. Specifically, we include the first two principal components as covariates both in the expert network of Eq. (4) and in the gating network of Eq. (5), so that $\mathbf{x}_d=\mathbf{x}_d^*$, $\forall d$. Indeed, the inclusion of further principal components does not lead to marked improvements in terms of the popular leave-one-out cross-validation information criterion (LOOIC).

An exploratory hierarchical cluster analysis based on all the upazila-specific covariates was conducted. It confirmed that the optimal number of clusters is two and the identified clusters can be easily labelled as *rural* and *urban*; the first includes peri-urban, rural, and remote areas while the second, smaller in size, encompasses metropolitan and highly urbanised upazilas.

Posterior inference has been conducted using Markov Chain Monte Carlo (MCMC) methods. Specifically, the implementation involved the utilization of the Hamiltonian Monte Carlo sampling algorithm through the `Stan` language and the `rstan` package. The estimation was carried out using four chains, each comprising 6,000 iterations, where the initial 3,000 iterations were designated as warm-up and subsequently discarded. Concerning the FH-MoE model, label switching problems are avoided as cluster-specific intercepts β_{01} and β_{02} are constrained to be ordered, namely $\beta_{01} < \beta_{02}$. In any case, this issue would not impact θ_d since the linear combination of the two components retains its significance.

The results obtained under the proposed FH-MoE model are compared against the FH model (other comparisons are included in an extended version of this manuscript). FH-MoE is clearly better than FH in terms of LOOIC: 134.3 (28.7) versus 186.6 (35.6). The reason is apparent from Table 1, where we can note that the posterior means of regression parameters are different. Namely, the intercepts are widely different and so are the slopes associated to a lesser extent those associated to the second component. Credible intervals have different sizes in the two groups, because of their different size and dispersion.

Table 1. Posterior means and 90% credible intervals related to the basic parameters which rule the models at the linking levels

	FH		FH-MoE			
Comp.	Unique		Cluster 1		Cluster 2	
Par.	Est	90% C.I.	Est	90% C.I.	Est	90% C.I.
β_0	–0.08	[–0.11,–0.04]	–0.54	[–0.82,–0.32]	0.49	[0.20,0.99]
β_1	0.66	[0.62,0.70]	0.42	[0.10,0.73]	0.42	[0.23,0.55]
β_2	0.07	[0.04,0.11]	0.01	[–0.05,0.07]	0.13	[–0.07,0.32]
σ_u	0.33	[0.30,0.36]	0.11	[0.03,0.17]	0.52	[0.41,0.69]

More importantly, the posterior of component-specific variance components σ_u^2 are also different, reflecting the different explanatory power of the auxiliary information in the two clusters and namely, the better fit in cluster 1. This implies that mixture-of-experts methodology enables a more pronounced shrinkage in the rural cluster compared to the urban cluster, addressing shortcomings of a standard Fay-Herriot model which would imply uniform shrinkage across areas with differing characteristics. Additionally, the synthetic components to which we shrink differ between clusters due to variations in the relationship between auxiliary variables and the target variable. The different shrinkage processes allowed by FH-MoE with respect to the FH are apparent in Fig. 1.

The presence of two regimes distinguished by the FH-MoE leads to notable improvements also in the estimates for OOS areas, as we can exploit the mixture to assign the area for which we do not have observations to its most likely group.

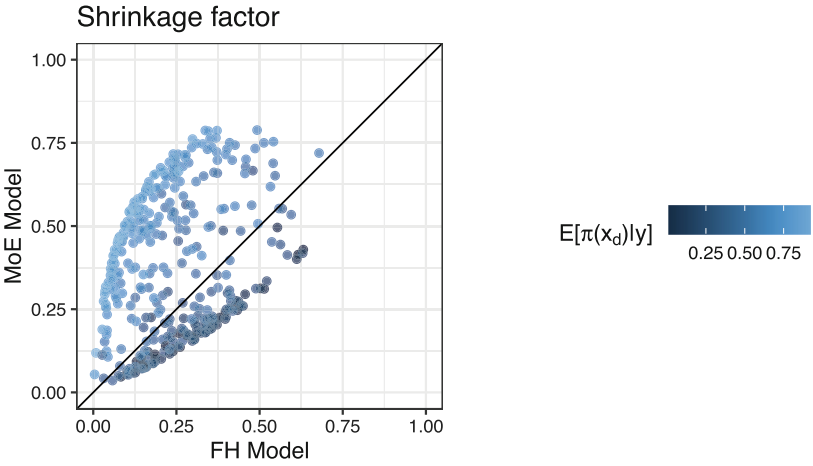


Fig. 1. Shrinkage factors associated to area level estimates computed according to FH and FH-MoE models

References

1. De Nicoló, S., Fabrizi, E., Gardini, A. Extended beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh. Ann. Appl. Stat. **18**, 3229–3252 (2024)
2. Fay, R.E., Herriot, R.A.: Estimates of income for small places: an application of James-Stein procedures to census data. J. Am. Stat. Assoc. **74**, 269–277 (1979)
3. Fruhwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of Mixture Analysis. CRC Press, Boca Raton (2019)
4. Goodrich, B., Gabry, J., Ali, I., Brilleman, S.: `rstanarm`: bayesian applied regression modeling via Stan. R package version 2.1 (2020)
5. Rao, J.N.K., Molina, I.: Small Area Estimation. John Wiley & Sons, Hoboken (2015)
6. Gershunskaya, J., Savitsky, T.D.: Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. J. Surv. Stat. Methodol. **8**(2), 181–205 (2020)

The Design of an Energy Data Eco-System

An Analysis of User-Specific Information Needs and Data Integration Processes

Gianna Greca[1] and Alessandro Zeli[2(✉)]

[1] Istat – National Accounts, Rome, Italy
greca@istat.it
[2] Istat– DVSE, Rome, Italy
zeli@istat.it

Abstract. Over the last twenty years the energy sector has been characterized by numerous and important changes. The pursuit of new and ambitious energy and environmental objectives determines the onset of further changes and the need to have information bases to support the monitoring and development of policies. The current organization of the statistical offer of the energy sector within SISTAN is based on a division of labor between some highly specialized subjects in the production and analysis of data within their institutional competence. The exchange and integration of data, as well as the response to requests from international bodies (UN, OECD, IEA, Eurostat), are guaranteed by the network of SISTAN statistical offices. However, its further evolution with a view to creating a "data eco-system" will lead several advantages (expansion and differentiation of the output, increase in the coherence, efficiency and timeliness of the data produced, etc.). The energy eco-system data will be instrumental in bridging the information gap relating to the growing demand for statistical information for the achievement of energy and climate objectives and the satisfaction of the information needs of the PNRR, in promptly providing statistical information to policy makers, as well as correctly recording in National accounts support measures in the energy and environmental fields. For the first time, in this paper we propose the conceptual scheme, the methodological and statistical framework for the construction of the energy data eco-system with an output example on energy products.

.

Keywords: data eco-system · energy sector · data integration

1 Introduction

Over the last twenty years the energy sector has been characterized by numerous and important changes. The pursuit of new and ambitious energy and environmental objectives determines the onset of further changes and the need to have information bases to support the monitoring and development of policies. Istat is determined to contribute in facing the new needs in information at national and international level, in particular in

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 63–68, 2025.
https://doi.org/10.1007/978-3-031-64346-0_11

order to bridging the information gap relating to the achievement of energy and climate objectives and the satisfaction of the in-formation needs of the PNRR (*Piano Nazionale di Ripresa e Resilienza*), in promptly providing statistical information to policy makers, as well as correctly recording in National Accounts (NA) support measures in the energy and environmental fields. These targets may be realized through the enhancement of a deeper collaboration of all data-owner stakeholders. This collaboration has to be aimed to the building of an integrate database allowing an easy and timely analysis of energy sector trends. The first step completed by Istat is the constitution, in the framework of National Statistical Program (PSN), of a Focus Group aimed to the "integrated analysis of energy sector". The members of the Focus Group, among the most important official statistics suppliers, agreed to launch a proposal of a project named "The energy national system: towards an eco-system for energy statistical data". The project, approved as part of the 2024 update of the 2023-2025 PSN, collected the participation of Istat, Ministry for the environment and the energy security and many other entities operating in the energy field [1]. The project aim is to plan and realize an integrated system for the statistical data production and dissemination (statistical data eco-system) containing in a unique platform data coming from the Bodies participating to the SISTAN (organized in order to return the updated and validated data to stake-holders).

2 Conceptual Scheme

The notion of eco-system states that in every close system the participants have to cooperate to maintain the stability of the system itself and maximize the benefits for all the players involved. Therefore, every participant entity has to define new strategies and activities in order to support the services coproduction and management and, finally, promote a sustainable use of the common resources. The current organization of the energy statistical supply, in the SISTAN framework, is based on a division of labor between subjects highly specialized in the production and analysis of data of its own competency. The data exchange and integration, as well as the fulfilling to requests of international bodies (ONU, OCSE, IEA, Eurostat), are granted by the SISTAN statistical offices network. However, the evolution of this system towards a statistical data eco-system can determine a list of advantages such as: enlargement of output in terms of data granularity and new indicators productions without collection costs; increase of data consistency; increase of efficiency and timeliness in data production and integration; possibility to differentiate the output according to the stakeholders' information needs on the basis of shared quality standards and, finally, avoid duplication and unnecessary steps during the information transmission. The implementation of PNRR (including the REPowerEU project) and the new information needs make it necessary to go beyond the current system to make it possible the exploitation of the potential inherent in the availability of shared data in a single and common register. This new system, making possible the interoperability between the different databases allows the creation of value added in terms of information, in other words it allows the identification of latent relationship and variables resulting from the integration of sources [2]. In this way we can overcome the current situations of "non-integration" (according to the possible availability of databases) or of "vertical integration" and move to an "eco-system" of data.

The new information system can be modelled on ISTAT's Integrate Registers System (IRS). The IRS system have to be re-modelled in order to take in account of the participation of different entities having different specificity and goals. In particular, the information system has to respect some criteria: the partnership between stakeholders, the accessibility and the safety of the shared data. Hence, in this first phase of the eco-system development the data management is structured in a dual configuration: a "core" structure, namely a structure within which an entity or an IT function (in this case the eco-system) plays as a pivot but maintaining, for the single entity, the possibility of an autonomous action in the data management and dissemination (Fig. 1). A centralized data management in a statistical information system implies a system connecting the statistical data sources and integrating the sharing and utilization modes of the basic statistical information elaborated by stakeholders involved. This permits to open a multilevel access to the whole set of shared data. In a data governance model with an "open" data eco-system: the technological, organizational and functional solutions are defined with respect to the whole data set and so "optimized" at the system level; the adopted solution in terms of information safety and data confidentiality are aligned with the best practices; the investment in data quality shall be agreed between stakeholder and are aligned with international standards.

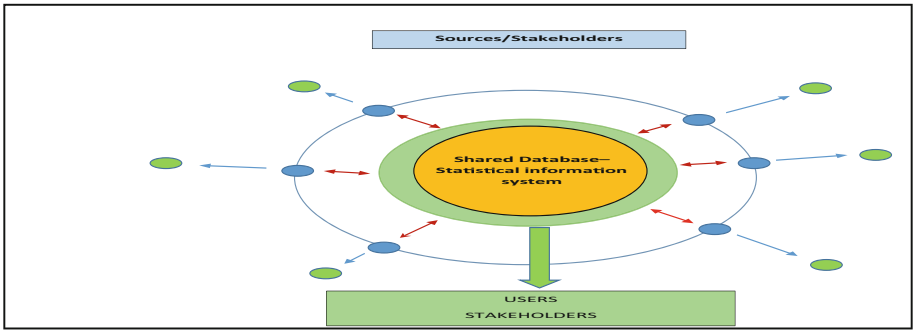


Fig. 1. The information integration system as final goal of the project

Several measures should be taken to achieve these objectives. In particular, checks should be carried out to ensure consistency of data from different sources, first at a logical level and then at a statistical level. At this point of the logical-physical building of the database, the most relevant information, by a statistical point of view, have to be identified. Further checks on statistical quality, consistency and compliance with the definition requirements (at the level of EU regulations) will be carried out on this subset of information. For this sub-set of data could be prepared ad hoc data extraction system, for instance in order to prepare periodical dissemination notes or publications. The database will provide IT safety checks, in this case should be planned secure "access points" for all authorized stakeholders with appropriate protocols (Fig. 2).

The project implementation provides for several steps. In the preliminary step, a survey (mapping) of information needs will be carried out. In the first step a technical verification of the degree of interoperability of the data and the effective statistical

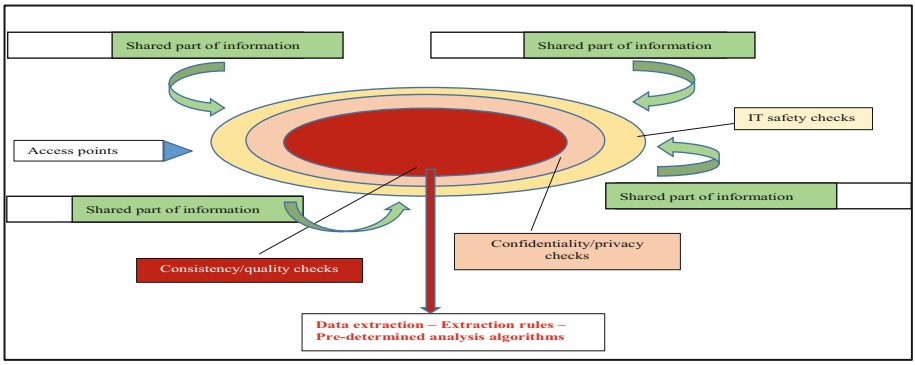


Fig. 2. Concentric check levels – overall scheme

integrability (consistency, quality, etc.) will be carried out. The second step will prepare the conceptual and functional design of the ecosystem. In the third steps will be lunched the experimental integration of data. In the last steps the system and the dissemination plan will be up and running [5].

3 Methodological and Statistical Framework for the Construction of the Energy Data Eco-System

The pursuit of new and ambitious energy and environmental objectives determines the onset of further changes and the need to have information bases to support the monitoring and development of policies. The creation and use of an integrated and shared data system will make it possible to satisfy various information needs more easily and more immediately. Furthermore, the integration of data of individual entities belonging to SISTAN into a single platform will make it possible getting updated and validated data available to *stakeholders*. To this aim, the integration of existing data sources is crucial in order to get an output based on the "supply-demand" scheme, in coherence with both the international Energy Balance scheme and the NA supply-use tables scheme [1–3]. This scheme typically provides, on the supply side, the production and import aggregates and, on the demand side, the aggregates relating to consumption by enterprises, families and the Public Administration and aggregates relating to exports and change in stocks. The choice of such scheme guarantees the consistency of definitions, classifications and registrations adopted at national and international level in the field of Energy Statistics and National Accounts. It will therefore be possible to build aggregates and indicators that can be immediately used by *policy makers*. The innovative idea of building an energy data eco-system is to move from the "data silos" of each body or entity entitled to data collection to the creation of "integrated information", with the aim of providing an "Information Pyramid" with three different output levels (Fig. 3) [1–3]. The first level, "Basic data", involves the creation and making available of micro data on energy and on the main aspects of the economy related to energy. The main characteristic of the output at the first level of the pyramid is that of being able to have information on either i) individual energy products; ii) production units; iii) economic-legal units involved in

the production, distribution and consumption of energy; iv) universe of units statistics or representative samples of the individual energy sector supply chains (whether or not belonging to regulated markets); v) representative samples of families; vi) main lists of Public Administration units. The integration among sources will also make information available at territorial and size class of employees level of detail.

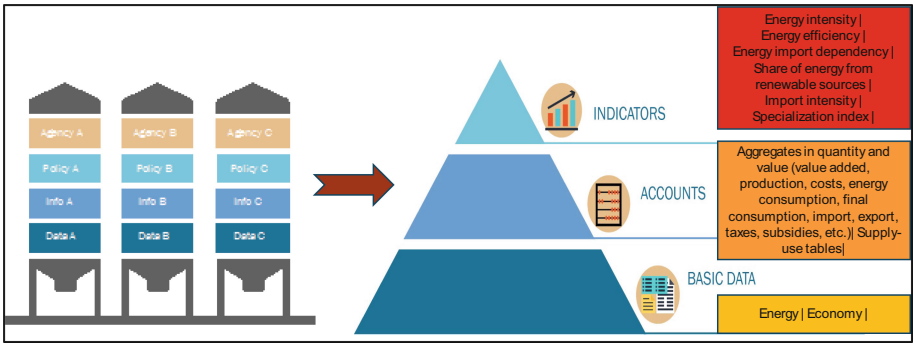


Fig. 3. Energy data ecosystem information pyramid. Source: [4]

The second level of the pyramid offers greater information power than the previous one. The primary objective is in fact the estimate and processing of "Accounts/Balances/Aggregates", in particular of specific aggregates in quantity and/or value such as production, costs, value added, imports, exports, intermediate consumption, final consumption, etc. The underlying idea is, as mentioned, the integration of data sources and the building of an output based on the "supply-demand" scheme, in coherence with both the international Energy Balance scheme and NA supply-use table scheme. The third level involves the development of indicators. It is necessary to combine the analysis of the aggregates and/or specific structural variables (number of units, etc.) with specific indicators representative of the energy performances and economic results at the level of the entire economy, single sector of economic activity and individual economic operators, carrying out comparisons of positioning at international level. Below are some *key indicators* that can be built from scratch or whose informative scope can be increased: *i)* Energy intensity; *ii)* Energy efficiency; *iii)* Energy dependence on imports; *iv)* Share of energy from renewable sources); *v)* Intensity of imports; *vi)* Specialization index; *vii)* Concentration ratio; *viii)* Vertical integration; *ix)* Cost competitiveness; *x)* Energy poverty of families and businesses. The system must also ensure the possibility of monitoring the evolution of phenomena over time, especially in relation to those characterized by greater dynamism, guaranteeing information continuity. Finally, in addition to elementary and aggregate data, the eco-system will contain metadata (internationally standardized classifications) to support the conceptual and functional integration of the plurality of information sources used. Consistently with the "supply-demand" scheme described above, Fig. 4 shows an example of output expressed in economic terms with references to oil and natural gas [2, 3].

In conclusion, the energy data eco-system allows: a leap in the concept of integration both with reference to the sources and with reference to the *governance* model

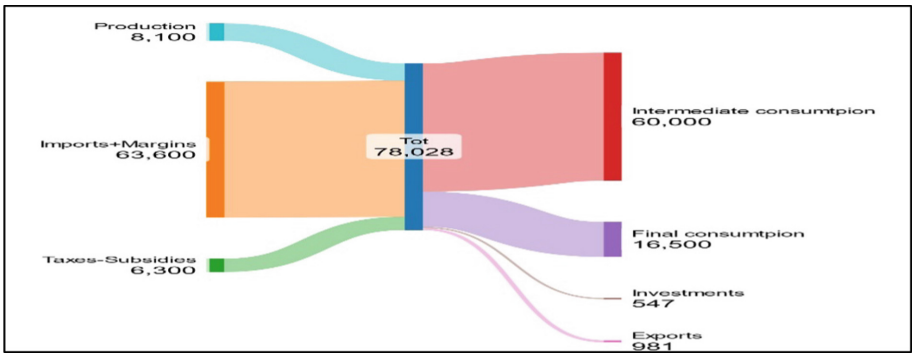


Fig. 4. Output example: economic flows of "oil and natural gas" (*thousands of euros*) Source: [3] – 2020 Istat data

(eco-system versus stovepipe model); a qualitative leap in the construction of a system defined on a "supply and demand" scheme in coherence with both international Energy Balance scheme and NA supply-use tables scheme; a leap in level and innovation in the construction of a system which, through the "Information Pyramid", provides not only an integration between microdata, but the processing and construction of aggregates and indicators that can be immediately used for the purposes of analysis and monitoring of specific policies. Moreover, it permits a qualitative leap in the construction of a system that returns, for the first time, economic information, integrating and harmonizing the information coming from different data sources.

References

1. Greca, G.: Proposta Studio progettuale (Stu) sull'energia, PSN 2023–2025 - Aggiornamento 2024–2025, Focus group analisi integrata del settore energetico, Istat, 3 aprile 2023 (2023)
2. Greca, G., Zeli, A.: Linee guida per la progettazione di un eco-sistema di dati sull'energia e sulle fonti rinnovabili, Documento progettuale, Istat, settembre 2023 (2023)
3. Greca, G.: Illustrazione dello schema concettuale di riferimento dell'eco-sistema dei dati sull'energia: output e principali vantaggi informativi per gli stakeholders- - Incontro tecnico "Eco-sistemi di dati sull'energia", Istat, 14 dicembre 2023 (2023)
4. Servizio studi della Camera dei Deputati: Documentazione di finanza pubblica n. 28/R/1 - Le proposte del Governo per la revisione del PNRR e il capitolo REPOWEREU, (2023)
5. Zeli, A.: Illustrazione delle principali caratteristiche tecniche e funzionali di un eco-sistema di dati sull'energia e macro-fasi del progetto di sviluppo - Incontro tecnico "Eco-sistemi di dati sull'energia", Istat, 14 dicembre 2023 (2023)

Generation of Synthetic Data from Mobile Network Operators (MNO) Data Through Generative Adversarial Networks (GANs)

Francesco Pugliese(✉), Angela Pappagallo, and Massimo De Cubellis

Italian National Institute of Statistics, 00184 Rome, Italy
frpuglie@istat.it

Abstract. In recent years, institutions have explored synthetic data generation to retain statistical properties and enhance privacy. Mobile network data (MND), containing sensitive call records and SMS activity, exemplifies this need. Generative Adversarial Networks (GANs) are crucial for synthesizing tabular data like MND, with Copula Conditional Tabular GANs (Copula GANs) standing out for preserving spatial distribution and privacy. However, replicating the temporal structure of call couplings may require advanced models like Time GANs or Graph GANs, employing Recurrent Neural Networks (RNNs) or Transformers. These models enhance the replication of bivariate joint distributions by handling temporal dependencies. The study emphasizes Copula GANs' promise for synthetic data generation and suggests exploring advanced GAN topologies for improved temporal structure replication.

Keywords: Synthetic Data · GAN · Machine Learning · Artificial Intelligence

1 Introduction

Public and private research institutions have explored various methods for generating synthetic data for statistical purposes, maintaining similar characteristics to real datasets while enhancing privacy [1]. Synthetic data is immediately usable for computing confidential statistics, providing a key advantage over real data. Mobile network data (MND) is often used for such analysis, including call records and SMS activity, considered private by providers [2, 3]. Statistical Institutes adopt a process pipeline approach, similar to the 'ESSNET Big Data' project, utilizing raw data computation by providers like Mobile Network Operators [4]. Employing Generative Adversarial Networks (GANs) [5], specifically 'CTGAN' [6] synthetic data is generated from telephony data, assessed using utility and privacy metrics from the 'SDGym' framework.

2 Methods

2.1 Generative Modeling

Generative Modeling employs unsupervised learning to uncover patterns in existing data, generating new synthetic data. Generative Adversarial Networks (GANs) operate with a generator (G) and discriminator (D), distinguishing between real and synthetic data.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 69–74, 2025.
https://doi.org/10.1007/978-3-031-64346-0_12

During training, the generator maximizes the discriminator's error probability, creating a minimax two-player game. Copula GANs merge Gaussian Copulas with CTGANs, addressing challenges in tabular data. They handle mode-specific normalization and efficiently preprocess data, crucial in higher dimensions. Copula CTGANs, proposed for training and synthesizing data, leverage SDGym, enhancing generation of tabular data with diverse types and distributions, ensuring stable and efficient training.

2.2 Experimental Setup

The process of generating synthetic data involves the use of Mobile Network Operator data with four variables: SIM code, antenna/sector code, time, and Call Detail Records (CDR). The dataset comprises 10,000 CDRs, which provide details of telephone calls passing through telecommunication equipment. The attributes include caller identification (SIM code), temporal information, and call origin (Cell_Call_Code). To ensure privacy, the data is pseudonymized and standardized attribute data types are used. Symbolic attributes are transformed into categorical variables, while others represent continuous variables for call date and time. 'TIME_MIN_CALL' combines 'Call Time' and 'Call Date', simpli-fying features. 'FESTIVE' identifies bank holidays or weekends. Renaming attributes aids the Generative Adversarial Network (GAN) in capturing spatio-temporal relation-ships between 'NUM_CALLER_KEY' and 'CELL_CALL_CODE'.

3 Results

Utilizing the capabilities of Copula GAN and the Synthetic Data Gym Framework, we've generated 10,000 synthetic telephony data points, matching the quantity of the actual Mobile Network Operator (MNO) data. Our primary objective remains maintaining the integrity of the data structure and statistical properties of the original dataset. The significance of this study lies not only in its innovative use of Generative Adversarial Networks (GANs) in Official Statistics, but also in the extensive exploratory, descriptive, and visual analyses conducted to assess the utility and privacy aspects of the newly generated synthetic data.

3.1 Univariate Analysis

The value ranges of the two key synthesis variables, CELL_CALL_CODE and NUM_CALLER_KEY, remain consistent with those in the original dataset. Our synthesis process involves appropriately combining these variables to capture their relationships effectively. Hence, in this univariate analysis, we've tallied the occurrences of each value for these variables in both the original Mobile Network Operator (MNO) dataset and the synthesized one. The image depicts similar distributions of CELL_CALL_CODE and NUM_CALLER_KEY variables, suggesting successful replication by our synthetic generation model. Moving to Fig. 1, a univariate Kernel Density Estimation Analysis (KDE) focuses solely on these variables, estimating the Probability Density Function

(PDF). KDE provides a smoother representation compared to bar plots, enhancing interpretability. The distributions of both original and synthetic data show comparable ranges, densities, shapes, and patterns. This indicates that the Copula GAN model effectively captures the univariate characteristics of the original data, successfully replicating them in the synthetic data. Various frameworks were experimented with in this study besides SDGym and found that reproducing these distributions on this type of data is challenging. Hence, we are highly satisfied with the outcomes achieved.

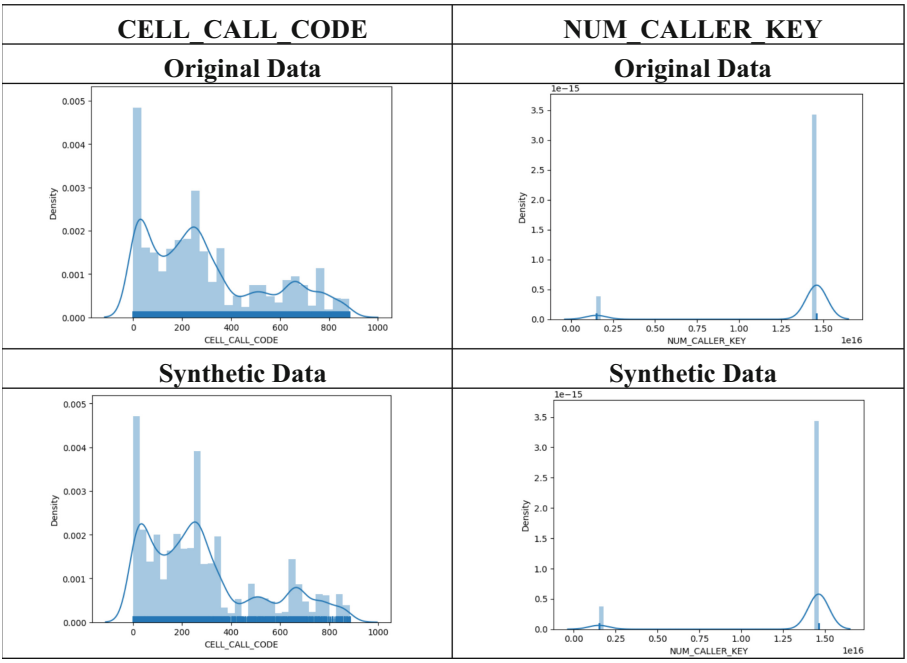


Fig. 1. Plot of Univariate Distrubution the Original and Synthetic by Kernel Density Estimation.

3.2 Bivariate Analysis

This analysis compares the joint distributions of CELL_CALL_CODE and NUM_CALLER_KEY in the Original and Synthetic datasets, categorized by the engineered variable 'FESTIVE'. A scatter plot (Fig. 2, left) reveals a subtle nonlinear correlation in the original data that is absent in the synthetic plot. This indicates that the synthetic model has not captured the intricate spatio-temporal relationships. The complexity involves discerning user habits and cell interactions. Future improvements may require the exploration of different GAN topologies. Furthermore, a multivariate Kernel Density Estimation chart (Fig. 2, right) illustrates the differences in joint distribution between the Original and Synthetic datasets, providing insights for further refinement.

We can see the same density distributions of PDF in both datasets: Original and Synthetic, which give a clue that this property is preserved in the synthetic output data.

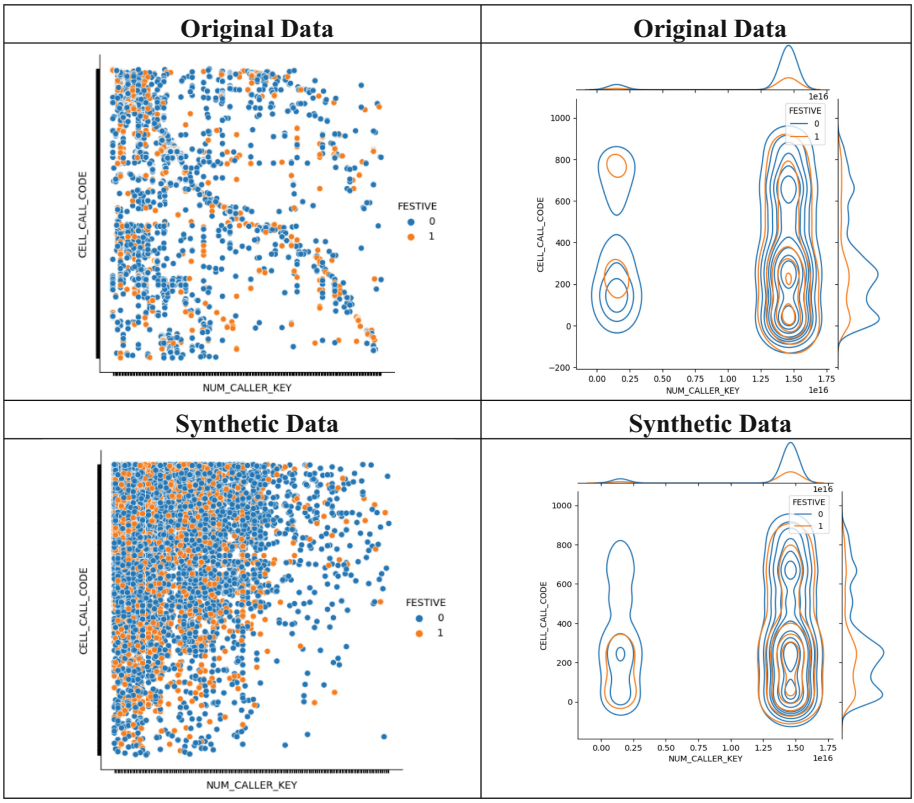


Fig. 2. Plot of Joint Distrubutions and KDE of the Original and Synthetic of main features.

3.3 Aggregate Analysis

This study compares aggregated indicators, such as means and standard deviations, for critical features, including CELL_CALL_CODE, NUM_CALLER_KEY, and FESTIVE, in both the original and synthetic datasets. Additionally, Cumulative Sums Distributions for each feature are analyzed to depict the probability of random variables, aiding in understanding the dataset's characteristics.

In Fig. 3 instead, we have the cumulative sums per feature of the 2 distributions: original and synthetic (real and fake). As we can observe there is a very high overlapping among the distributions points of original and synthetic datasets.

3.4 Machine Learning Analysis

Various Machine Learning Classifiers were used in this study to train on a specific target variable, FESTIVE, using other features as input. The performance was compared between real and synthetic datasets, aiming for similar results. The F1-score, which balances precision and recall, was selected due to potential class imbalance. The results (see Table 1) showed identical F1-scores for real and synthetic data, indicating similar

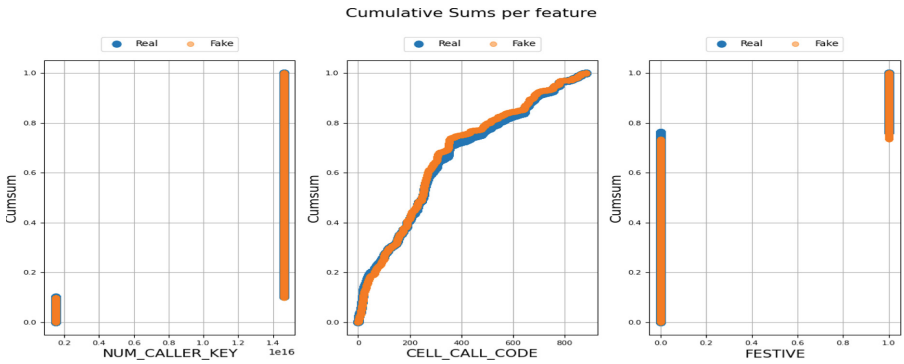


Fig. 3. Cumulative Sums per feature of Original and Synthetic Data.

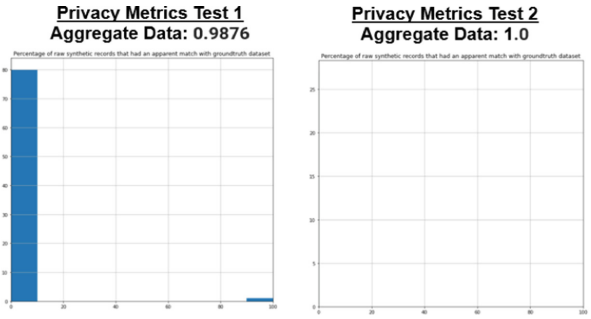
statistical properties. Synthetic data appears to accurately replicate the statistical characteristics of the original dataset, as demonstrated by the performance of the machine learning classifiers.

Table 1. Machine Learning classiefiers Metrics.

Classifier	F1-Score
DecisionTreeClassifier_fake	0.6240
DecisionTreeClassifier_real	0.7650
LogisticRegression_fake	0.7485
LogisticRegression_real	0.7575
MLPClassifier_fake	0.2515
MLPClassifier_real	0.2425
RandomForestClassifier_fake	0.6485
RandomForestClassifier_real	0.7600

3.5 Privacy Analysis

Privacy metrics are used alongside utility metrics to assess the quality of synthetic data, although there are fewer options available. Two privacy metrics tests were conducted. Test 1 showed a slight privacy failure when merging based on ANTENNA_CODE, while Test 2, which merged based on SIM_CODE, showed no failures. These tests evaluate matching patterns between synthetic and original datasets. The privacy levels are synthesized into an Aggregate Privacy Metric (APM) that ranges from 0 to 1. A high APM indicates strong privacy preservation. The test results yielded 0.9876 and 1.0, indicating very high to maximum privacy, respectively (see Fig. 4).



$$\mathbf{APM} = \mathbf{1} - \frac{\sum_{k=0}^{n} \frac{S_k}{100}}{N}$$

Fig. 4. Original and Synthetic Privacy Metrics Formula (APM) and Plots

4 Conclusions

Our experimental setup shows that Copula Conditional Tabular GANs (Copula GANs) are excellent at generating synthetic data while preserving spatial distribution and high privacy levels. Comparisons with other GAN models, including those built from scratch, have revealed that Copula GANs have superior synthesis capabilities within SDGym. They effectively reproduce univariate distributions and statistical properties of the original data. However, replicating the temporal structure of all call couplings may require more advanced models such as Time GANs or Graph GANs. Time GANs use Recurrent Neural Networks (RNNs) to handle temporal dependencies, while Graph GANs learn ontology structures to capture temporal relationships, improving the replication of bivariate joint distributions.

References

1. Soltana, G., Sabetzadeh, M., Briand, L.C.: Synthetic data generation for statistical testing. In: 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2017
2. Xu, F., et al.: Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In: Proceedings of the 26th International Conference on World Wide Web (2017)
3. Ricciato, F., et al.: A reflection on privacy and data confidentiality in Official Statistics. ISI (2019)
4. Ricciato, F., Grazzini, J., Museux, J.M.: Public manuals and open-source code: rethinking methodological documentation for new data sources, NTTS 2021 (2021)
5. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
6. Xu, L., et al.: Modeling tabular data using conditional gan. Adv. Neural Inf. Process. Syst. **32** (2019)

Data Ecosystems and the Measurement of Added Value from Data Integration: A Methodological Framework

Paolo Righi(✉)

Italian National Institute of Statistics – Istat, Roma Via Balbo 16, Rome, Italy
paolo.righi@istat.it

Abstract. Data Ecosystem (DE) represents a new production model that is increasingly adopted by public and private stakeholders to share investment costs and maximize data sharing. While the benefits and costs associated with a DE are clearly perceived, its accurate definition both in terms of scope of activity, mechanisms and incentives for operation have not yet been sufficiently explored.

This paper seeks to define a theoretical and conceptual framework within which DE can be properly analysed. In particular, the methodological approach to measure the information return of the data integration process also in economic terms is identified in the theory of the Value of Information (VoI) approach. This approach, based on Bayesian statistical inference, assumes the benefits of a decision-making have a direct relation with the quantity and the quality of available data. The paper outlines the possibility of extending the approach to the specific case of DE, highlighting its advantages and potential problems.

Keywords: data integration · data quality · Data Ecosystem · Value of Information (VoI)

1 Introduction

Data Ecosystems (DEs) have recently emerged as a new form of data production that can increase the possibility of integrating new data and improving the quality of already integrated data. Despite their growing relevance, their specificity and operating mechanisms have not been sufficiently explored.

The paper outlines a conceptual and methodological framework within which DEs can be properly analysed, both theoretically and empirically. The second section highlights the main characteristics of DEs compared to other data production models, as well as the main drivers of a successful DE. Since the economic return generated by data sharing plays a crucial role in the setting up and maintaining a DE, the third section describes a methodological framework for the measurement of the information added value generated by data integration. Section four introduces the VoI approach in the DE context. Section five gives some conclusions.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 75–80, 2025.
https://doi.org/10.1007/978-3-031-64346-0_13

2 The Key Characteristics of a DE

Data sharing, along with data integration, is one of the key drivers of the current digital revolution. The ability to make the most of data integration depends on both methodological-technological and regulatory issues, but also on the choice of data production model. DEs are emerging as an alternative data production model with respect to both vertical data integration and open data (Moore, 1993; Olivera et. al. 2018, 2019; Altendeitering and Guggenberger, 2020).

The three production models have different characteristics, which give advantages and disadvantages when performing a statistical analysis (Table 1).

Table 1. Key characteristics of different data production models

Characteristics	Production model		
	Vertical integration	Open data	Data Ecosystem (DE)
Data standardization	High	Marginal	High
Data integration	Limited	High	High depending on incentives
Data quality	High	Marginal	High depending on incentives
Data platform and business analytics	Medium	Low	High

The vertical integration model maximises the benefits of high levels of data standardisation and quality, but it limits the data integration to its own information. The open data production model maximises the benefits of integrating freely available data, but is penalised by limited investment in data standardisation and quality. The DE production model is in an intermediate position. Comparing the vertical integration and DE production models, the latter offers greater opportunities for data integration as it simultaneously takes into account a larger number of data owners. On the other hand, given the right incentives, DE can make significant investments in data quality and in extending integration processes.

To explore the nature and function of a DE in more detail, it is useful to consider three key actors of a DE, as illustrated in Fig. 1 (left). At the highest level are the stakeholders of the project, that in general have a role of decision makers. Usually the stakeholders are data owners of their own data and the subjects who express the demand for information within the perimeter of the DE. The use of the DE by third parties (wholly or partly public) may also be envisaged. The information needs of the stakeholders are met by the data managers, which allows both the secure and efficient management of the integrated database in compliance with legislation (privacy and confidentiality) and the development of quality control and analysis tools. At the base of the pyramid is the level of proprietary control of the databases, which remain physically under the control of the individual data owners.

In a DE there are stakeholders of different nature who maximize their information gains while minimizing the costs deriving from having to face shared investments in

the design and maintenance of the data infrastructure. Figure 1 (right) shows three main levels.

Fig. 1. Main actors of a DE (left) and main levels of analysis of a DE (right).

At the bottom, the Legal-Institutional level refers both to the regulatory framework at national and international context and to the specific forms of contract between the actors. It defines the scope and methods of the integration processes. The economic behaviour of the key actors level refers to the behaviour adopted by the stakeholders, with particular attention to the cost-benefit trade-off to increase the level of cooperation and economic investment in the DE. At the top, the Methodological data management level concerns the set of technologies, methodologies and tools aimed at extracting information value for each stakeholder. Note that, the cooperation to the DE can be the only feasible strategy to access to new data not obtainable otherwise.

3 A Methodological Framework to Measure the Economic Value of Data Integration

More available data does not necessarily lead to a proportional increase in the added value of the information. Furthermore, the certain costs of data integration and data quality improvement must be outweighed by uncertain expected benefits. Although these sentences are self evident, their quantitative assessment is still an open issue in the scientific literature and business practice.

We focuses on the approach known in the literature as Value of Information (VoI) theory (Raiffa H, Schlaifer, 1961; Parmigiani, G.,Inoue, 2009) to provides a quantitative framework for quantifying the value of acquiring additional information based on the its contribution to reduce uncertainty in decision making. The VoI approach uses the Bayesian inference to a data selection/prioritization process with respect to a set of alternative decisions (ISPOR Report, 2020; Jackson et al., 2022).

The VoI approach usually exploits a decision analytic model describing the relationship between outputs relevant for decision making (e.g., costs and benefit) and the model input parameters corresponding to unknown population quantities. Observational studies, registries, or expert opinion provide imprecise estimates of these quantities. To keep the thing simple, let us assume a single discrete parameter defining $\theta = \left(\vartheta_1, \ldots, \vartheta_j, \ldots, \vartheta_J\right)'$ the vector of possible values of the parameter being ϑ_j a scalar. In general θ could be a continuous variables. Trivially, θ could represent the state of economic, social or environmental system. In general, the parameters are used within

a model used to mimic the real economic, social or environmental system. The uncertainty about the true unknown parameter value is expressed by a probability distribution, $p(\theta) = \{p(\vartheta_1), \ldots, p(\vartheta_j), \ldots, p(\vartheta_J)\}$ with $p(\vartheta_j) > 0$ for $j \in \{1, \ldots, j, \ldots, J\}$ being $\sum_j p(\vartheta_j) = 1$. We denote $p(\theta)$ as prior distribution before any planned future data integration. Let $D = \{d_1, \ldots, d_i, \ldots, d_I\}$ be the set of mutually exclusive decision options for a single stakeholder (for instance a business man, a corporation, a bank or a public institution), hereinafter denoted as decision maker. The decision model gives the outcome value (or utility), $V(d_i, \vartheta_j)$, when taking the decision d_i given the parameter ϑ_j. The $V(\cdot)$ function may also include as an exogenous information a price vector which represents the economic value of data based on shadow prices (willingness to pay). The uncertainty on the true ϑ_j allows to compute the expected value $E_\theta[V(d_i, \theta)] = \sum_j V(d_i, \vartheta_j)p(\vartheta_j)$ for the decision d_i. A risk-neutral decision-maker (Arrow and Lind, 1970) will choose the decision d_i such as $E_\theta[V(d_i, \theta)]$ is the $max_{d \in D}(E_\theta[V(d, \theta)]) = max\{E_\theta[V(d_1, \theta)], \ldots, E_\theta[V(d_i, \theta)], \ldots, E_\theta[V(d_I, \theta)]\}$ with $max\{\cdot\}$ the maximum value of the set. To better understand the VoI approach, the expected value with certainty is given according to the following reasoning: for each value ϑ_j the decision maker keeps the d_i such that $V(d_i, \vartheta_j)$ is the $max_{d \in D}([V(d, \theta)]) = max\{V(d_1, \vartheta_j), \ldots, V(d_i, \vartheta_j), \ldots, V(d_I, \vartheta_j)\}$. Again, before having the information giving the certainty of the true parameter, the decision maker can compute the expected value with certainty, $E_\theta[max_{d \in D}[V(d, \theta)]] = \sum_j max_{d \in D}[V(d, \vartheta_j)]p(\vartheta_j)$.

The difference of the two expected values gives a first index of the VoI denoted as Expected Value of Perfect Information (*EVPI*),

$$EVPI = E_\theta[max_{d \in D}[V(d, \theta)]] - max_{d \in D}(E_\theta[V(d, \theta)])$$

The *EVPI* is the expected benefit of eliminating uncertainty entirely. In general, the VoI represents the difference between the expected outcomes when a decision is made only on the basis of the prior information, and when new information is gained (Yokota and Thompson, 2004). In a real context new data does not enable to eliminate the uncertainty. A second VoI measure is the Expected Value of Sample Information (*EVSI*). Let us assume new data are obtained, for example making a survey sample on the target population or integrating new data from other decision makers cooperating in the DE. The decision maker updates the prior probability distribution with the posterior probability distribution. Let Y be the new data set provided by the DE, the posterior distribution is denoted by $p(\theta|Y) = \{p(\vartheta_1|Y), \ldots, p(\vartheta_j|Y), \ldots, p(\vartheta_J|Y)\}$ being $p(\vartheta_j|Y)$ the probability ϑ_j is the true parameter after observing Y. A straight and easy example is the following: assume the unknown parameter is the mean of the target population on a target variable. The decision maker estimate using the new data, Y, the mean of the target variable, $m_Y = m(Y, s)$, where $m(\cdot)$ is a function (estimator) on the Y data and the feature of the process generating the data s (for example the sample design, or the selectivity of the DE integrated data). Then, the decision maker compute the $p(\vartheta_j|m(Y, s))$. Note that we replace Y with a low dimensional sufficient statistic, $m(Y, s)$ in the posterior distribution. The natural framework for achieving the posterior distribution is the Bayesian approach. The decision maker updates the expectation by $E_{\theta|m}[V(d_i, \theta)] = \sum_j V(d_i, \vartheta_j)p(\vartheta_j|m)$, with $m \equiv m(Y, s)$. The risk-neutral decision-maker will choose the decision d_i such that $E_{\theta|m}[V(d_i, \theta)]$ is the $max_{d \in D}(E_{\theta|m}[V(d, \theta)])max\{E_{\theta|m}[V(d_1, \theta)], \ldots, E_{\theta|m}[V(d_i, \theta)], \ldots, E_{\theta|m}[V(d_I, \theta)]\}$.

Again, the decision maker has to decide to collect new data, so the sufficient statistic, m, is not observed yet. The VoI process compute the probability distribution, $p(m) = \sum_j p(m|\vartheta_j)p(\vartheta_j)$, and finally obtain the expected value if observing new data, $E_m\left[max_{d\in D}\left(E_{\Theta|m}[V(d,\theta)]\right)\right] = \sum_{m\in M}\left[\sum_j V\left(d_i, \vartheta_j\right)p(\vartheta_j|m)\right]p(m)$, where M is the set of the possible value of m. A second index of the VoI is

$$EVSI = E_m\left[max_{d\in D}\left(E_{\Theta|m}[V(d,\theta)]\right)\right] - max_{d\in D}(E_\Theta[V(d,\theta)]).$$

The $EVSI$ has to be compared with cost of collecting the new data. The difference between expected benefit ($EVSI$) and the cost (C) gives the Expected Net Benefit (ENB). The $ENB = EVSI - C \geq 0$ is a necessary condition for integrating new data or improving their quality.

4 An Extended VoI Approach to Include DE

We outline the extension of the VoI approach to DE introduced in Sect. 2. In this case, multiple decision makers want to benefit from data sharing, while the fixed costs of integrating the data remain common to all stakeholders. The expected benefits of DE are specific to each stakeholder, but not independent. Indeed, the entry or exit of stakeholders who also act as data owners in the DE affects the benefits of other stakeholders, since the amount of information available changes. In the following, we consider a proposal to reformulate the VoI approach with respect to the first issue.

Let us assume there are more than two decision makers that want to benefit from DE ($H \geq 2$). Let us define the basic set-up: there are different target parameters. Let $\Theta = \left(\theta_1, \ldots, \theta_q, \ldots, \theta_Q\right)'$ be the vector of the interest parameters. Each decision maker is potentially interested to know the true values of all or some of these parameters. The decision models of the hth ($h = 1, \ldots, H$) decision maker, is denoted by $V_h\left(d_{ih}, \vartheta_{1j}, .., \vartheta_{qj}, \ldots, \vartheta_{Qj}\right)$, where: $\theta_q = \left(\vartheta_{q1}, \ldots, \vartheta_{qj}, \ldots, \vartheta_{qJ}\right)$ is the vector of the parameters; d_{ih} is the ith decision stakeholder specific; $V_h(\cdot)$ is the decision stakeholder specific, being $V_h(\cdot)$ defined by functional form including all or some of the θ_q elements. The cooperation to the DE of the hth decision maker depends on the $ENB_h = EVSI_h - C_h >= 0$ in which ENB_h, $EVSI_h$ and C_h are respectively the ENB, $EVSI$ and C specific of the hth decision maker. We assume the $V_h(\cdot)$'s are independent each other and do not change in the DE; the $EVSI_h$ uses $p_h(\theta)$ (the prior distribution varies among the decision makers), and $p_h(\theta|m) = p(\theta|m)$ in the DE (the posterior distribution are the same); $C_h = CF + CV_h$ in which CF is the fixed cost to be involved in DE and CV_h is the variable cost specific for each decision maker. In a fully collaborative DE the CV_h can be included in the CF. It is reasonable that DE offers the opportunity to acquire large new data sets at relatively low cost, and any decision maker will expect a large reduction in the uncertainty of the parameters and a large net benefit according to the VoI approach. In order to evolve over time by increasing investment in data quality and integration, additional costs must be offset by increasing benefits. The choice of the decision maker to enter or to further invest in the DE depends on the ENB_h, being $ENB_h \geq 0$ a realistic condition. Finally, a rough overall informative value of the DE is the global net benefit, $GENB = \sum_h ENB_h$.

5 Conclusion

The paper highlights the key features of a DE compared to other data production models and emphasises the importance of quantifying the expected information return in a DE. For this purpose, the value of information (VoI) approach is considered. The VoI approach integrates Bayesian inference and data prioritisation process with respect to a set of decisions. This VoI approach, originally conceived for a production model consisting of a single stakeholder, is here extended to DE with multiple stakeholders adopting mechanisms of cooperation but also opportunistic investment. The paper outlines the framework and provides some indications for further development. Future work will have to consider both the completion of the theoretical framework with a better modelling of the strategic interaction between stakeholders and a more rigorous analytical formalisation, also complemented by empirical examples.

References

Altendeitering, M., Guggenberger, T.: Data Quality in Data Ecosystems: Towards a Design Theory Conference Paper (2020)

Arrow, K.J., Lind, R.C.: Uncertainty and the evaluation of public investment decisions. Am. Econ. Rev. **60**(3), 364–378 (1970)

ISPOR Report: Value of Information Analytical Methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. Value of Health, 23(3), 277–286 (2020)

Jackson, C.H.: Value of information analysis in models to inform health policy. Annu. Rev. Stat. Appl **9**, 95–118 (2022)

Moore, J.: Predators and Prey: A New Ecology of Competition. Harvard Business Review. May–June 1993 issue (1993). https://hbr.org/1993/05/predators-and-prey-a-new-ecology-of-competition

Oliveira, M.I.S., Barros Lima, G.D.F., Farias Lóscio, B.: Investigations into Data Ecosystems: a systematic mapping study. Knowl. Inf. Syst. **61**(2), 589–630 (2019). https://doi.org/10.1007/s10115-018-1323-6

Oliveira, M.I.S., Lóscio, B.F.: What is a data ecosystem? In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209281.3209335

Parmigiani, G., Inoue, L.: Decision Theory: Principles and Approaches, Wiley, Chichester, UK (2009)

Raiffa, H., Schlaifer, R.: Applied Statistical Decision Theory. Harvard Univ. Press, Cambridge, MA (1961)

Yokota, F., Thompson, K.M.: Value of information literature analysis: a review of applications in health risk management. Med. Decis. Mak.Mak. **24**, 287–298 (2004)

Content-Based and Data-Driven Integration. Markov Chain Models for the Inspection of Response Dynamics in Psychological Testing

Andrea Bosco(✉)

Università degli Studi di Bari Aldo Moro, P. zza Umberto I, 1, Bari, Italy
andrea.bosco@uniba.it

Abstract. The integration of content-based and data-driven statistical analysis in psychological testing offers a powerful and refined approach, leveraging theoretical foundations alongside surface-level observations of individual responses. This synthesis enhances assessment reliability, fostering a more comprehensive understanding of human behavior and cognition.

Content-based analysis, rooted in psychological theories, allows for the systematic examination of individual responses, ensuring the design of theoretically sound measurement instruments. Conversely, data-driven analysis harnesses empirical evidence from test responses, uncovering patterns through techniques such as machine learning.

The method involved a random assignment of 200 students to fake good or honesty groups, each completing a 20-item questionnaire based on the Balance Inventory of Desirable Responding (BIDR). The present study explores response dynamics using Markov chains and likelihood log ratio (LLR). LLR discriminates between groups based on response sequences. Faking good participants showed response sequences consistent with the alternating pattern of positively and negatively keyed items more frequently than the group responding honestly.

In conclusion, integrating content-based and data-driven approaches in psychological testing yields a dynamic, adaptable process that is less biased in real-world applications. This methodology, exemplified through Markov chains and LLR, enhances assessment robustness and ensures adaptability and relevance in the evolving psychological landscape.

Keywords: Markov Chain models · Test response dynamics · Likelihood log ratio · Psychometrics

1 Introduction

1.1 Content-Based and Data-Driven Integration in Psychological Testing

The integration of content-based and data-driven statistical analysis in psychological testing represents a powerful and nuanced approach that combines the strengths of theoretical foundations with a more surface observation of individual responses. This synthesis

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 81–86, 2025.
https://doi.org/10.1007/978-3-031-64346-0_14

can enhance the reliability of assessments, promoting a potentially more comprehensive understanding of human behavior and cognition. Content-based analysis (Linn, 2011) involves the systematic examination of individual responses, according to psychological theories. On the other hand, data-driven analysis (Sprent, 2019) exploits empirical evidence derived from actual test responses. The integration of these two approaches is particularly beneficial in addressing the limitations of each separately. Content-based analysis ensures a conceptually sound - strategic - assessment. Simultaneously, data-driven analysis adds an empirical - tactic - approach to the test responses, allowing for continuous refinement and validation of the instruments under examination.

For a long time, social sciences emphasize the dynamic interaction between participants and survey or questionnaire (e.g., Schuman & Presser, 1996; Schwarz & Sudman, 2012). The coherence in this process goes beyond question content, acknowledging the impact of past experiences (e.g., item order effect) on responses and introducing the concept of path dependency where past choices shape forthcoming outcomes. In the legal sciences positive *autocorrelation* is acknowledged between consecutive independent verdicts of the same jury (Bindler & Hjalmarsson, 2019). A Study by Du and Clark (2017) indicate the effectiveness of *autocorrelation* in explaining cognitive organization through first-order contingencies. Thiel and colleagues (2014) suggest the influence of *hysteresis* in a two-point discrimination test, asserting its importance when explicit stimulus properties are inconclusive. Hysteresis also emerges in a study of Odic, Hock and Halberda (2014) on numerical discrimination in children. In summary, path dependency, first-order autocorrelation, state-dependency, and hysteresis emerged in shaping responses to both simple and complex questions in various contexts. A further exploration into the foundational role of earlier responses in tests, surveys, and questionnaires is gathered by a 2012 paper. The authors Atmanspacher and Römer proposed a model based on *non-commuting observables* borrowed from quantum physics. This model posits that psychological processes measured with surveys or tests generate observable modifications like the order effects in experimental psychology (e.g., Hogarth & Einhorn, 1992). Markov chain models for analyzing sequential data seemed suitable methods to capture the influence of the most recent item on subsequent responses in psychological testing.

1.2 Markov Chains

Markov chains (e.g., Stewart, 2000; Strang, 2022) are a mathematical model that describe a sequence of events as a series of transitions between states. Following the *Markov property,* the probability of transitioning from one state to another depends uniquely on the previous one. The most relevant concepts in a Markov chain are:

State: a Markov chain consists of a set of distinct states that the system can occupy. These states represent different conditions. In the present study the space of the states is as follow:

$$\mathrm{S} = \{\text{Strongly disagree},\ \text{Disagree},\ \text{Neutral},\ \text{Agree},\ \text{Strongly agree}\} \tag{1}$$

corresponding to the response modalities. For the sake of simplicity each state is resembled in the subsequent tables with a number from 1 to 5, respectively.

$$\mathrm{S} = \{1, 2, 3, 4, 5\} \tag{2}$$

Transition: each state has associated probabilities for transitioning to other states. These transition probabilities define the likelihood of moving from one state to another in the next step.

Transition Matrix: the transition probabilities are organized into a transition matrix. Each element of the matrix represents the probability of transitioning from one state to another.

Initial Distribution: The chain also requires an initial probability distribution that describes the likelihood of the system starting in each state. In the present study the initial distribution is fixed:

$$\Pi = \{0, 0, 1, 0, 0\} \tag{3}$$

Chains with Discrete Time: Markov chains can operate in either discrete time or continuous time. In the present study the time is discrete, indeed transitions occur at distinct intervals, after each item.

Self-transition: in the context of a Markov chain the tendency of the system to remain in its current state, is reflected in the probabilities on the principal diagonal of the transition matrix.

Steady Distribution: it describes the long-term probabilities of being in each state. In a steady distribution, the probabilities do not change over time.

Moreover, the likelihood log ratio (LLR) is employed in the present study. The LLR is a statistical measure used to assess the probability distribution of a given sequence of responses and discern class membership based on at least two transition matrices coming from a theorical model or from data of other groups employed as training data. In the present study the sequences of two groups of participants instructed to respond to a psychological test by portraying a positive image of themselves (Fake good) or to answer honestly (Honesty) are processed according to the transition matrices obtained by the sequences of 12 experts in psychometrics and social desirability but completely unaware regarding the new approach of data analysis presented here.

Markov chains are widely used in various fields of science, to model systems that exhibit probabilistic and sequential behavior. To the best of our knowledge, they are not used previously detecting the dynamics of the response to a psychological test.

Finally, a description of the general hypotheses follows:

H1. Neutral Response (state 3): Fake good > Honesty group, comparing the two stationary distributions.

H2a. High adherence to a transition probability matrix resembling a drunkard's walk model (e.g., Nosofsky & Palmeri, 1997) of the Fake good with respect to the honesty group.

H2b. Alternatively, high adherence to test structure in Fake good with respect to the Honesty group, in line with the alternating pattern of positive and reverse keyed items (alternating pattern model).

H3. Positive LLR is associated to an indication that the observed sequence of responses is characteristic of the Gake good group (at the numerator), while a negative LLR is associated to the responses of the honesty group (at the denominator). The LLR is expected to be a robust estimator of the class membership.

2 Method

2.1 Participants

The study involved a sample of 200 university students. Participants were randomly assigned to two groups: one trained to provide socially desirable responses and the other encouraged to respond honestly.

2.2 Procedure

Prior to questionnaire administration, participants in the fake good group were instructed to present themselves favorably. The honesty group received no specific instructions regarding response presentation.

2.3 Materials

The survey instrument consisted of a 20-item questionnaire (from nr. 21 to 40 of the original version) dedicated to detects the *impression management* (the tendency to seek approval from others) as arranged in the Scale of the Balanced Inventory of Desirable Responding (BIDR) developed by Paulhus (1991) in which negatively and positively keyed items are completely alternated along the test. The reliability of the scale is good (alpha = 0.82). The study adhered to ethical guidelines. The research protocol was approved by the institutional ethics review board.

3 Results

H1. The steady state vectors for the two groups were calculated through the routines of the Python library *Numpy* (Harris et al., 2020). This method relies on the mathematical properties of eigenvectors and eigenvalues to find the steady state of a system represented by a transition matrix (e.g., Norris, 1998).

Faking good group steady state:

[0.22766083, 0.18964588, 0.09369275, 0.26010919, 0.22889135]

Honesty group steady state:

[0.12722855, 0.25029664, 0.14384796, 0.35118556, 0.12744129]

a) The probability in the neutral state is higher in the Honesty group.

b) The difference between the two steady state vectors is evaluated with the Kullback-Leibler Divergence (KLD) that appears to be very small (KLD = 0.0956). The H1 appears to be disconfirmed.

H2a. Divergence values between Fake good and Honesty groups with Drunkard's walk model matrices are large (KLD = 7.34 and KLD = 4.29, respectively). The H2a appears to be disconfirmed and, overall, the drunkard's walk model does not fit with the transition probability matrices of both two groups.

H2b. Divergence values between Fake good and Honesty groups with the alternating pattern model are small (KLD = 0.67 and KLD = 2.61). The H2a appears to be confirmed indeed the fake good group transition matrix fit better with the alternating pattern model than that of the other group. Overall, the alternating pattern model fit well with the transition probability matrices of two groups.

H3. Based on the LLR calculated by comparing the faking good and the honest matrices of the 12 experts, a score was obtained for each of the 200 participants (half instructed to fake good, the other half to respond truthfully). Additionally, for each participant, the BIDR impression management sub-scale scores were computed. This latter and the score obtained using Markov chains were employed in a study on Receiver-Operating Characteristics (ROC) for calculating the diagnostic power of the measures. The two indices – one content-based and the other data-driven – are effectively complementary. Table 1 reports the results.

Table 1. Results of the ROC study. Cut-off (>=): participants receiving a score equal or higher than that displayed are considered *positive*. Se: sensitivity, the proportion of true positives. Sp: specificity, the proportion of true negatives. Youden Index: it is calculated as the sum of Se and Sp minus 1, it succinctly captures the balance between sensitivity and specificity in a binary classification test. AUC: Area Under the curve, LR+: positive likelihood ratio, LR−: negative likelihood ratio

	Cut-off (>=)	Se	Sp	Youden Index	AUC	LR+	LR−
BIDR total score	68.00	0.39	0.99	0.38	0.68	39.00	0.62
LLR score	−1.33	0.91	0.83	0.74	0.88	5.35	0.11

So, a person with the probability equal to 0.3 of lying (e.g., Baer & Miller, 2002), a BIDR total score $>=$ 68 and a score $>=$ −1.33 on the LLR score, has a posterior probability of lying of approximately 99%. If he/she obtains a score $<$ 68 and an LLR $<$ −1.33 his/her posterior probability of lying drastically drops below 5% (calculations recurred to the application of the Bayesian Factor). Even in the scenario where the BIDR total score were to turn out negative and the LLR score, instead, is positive, the participant would still achieve almost a doubling of the prior probability (posterior = 0.59), laying the basis for further investigation. Therefore, the integration of these two methods derived from the same test appears very promising.

4 Conclusion

By combining content-based and data-driven approaches, psychological testing becomes a dynamic and evolving process. The integration of content-based and data-driven statistical analysis in psychological testing represents a harmonious approach that capitalizes on the strengths of both theoretical foundations and empirical evidence. This comprehensive methodology not only enhances the robustness of assessments but also ensures their adaptability and relevance in an ever-evolving psychological research.

References

Atmanspacher, H., Römer, H.: Order effects in sequential measurements of non-commuting psychological observables. J. Math. Psychol. **56**(4), 274–280 (2012)

Baer, R.A., Miller, J.: Underreporting of psychopathology on the MMPI-2: a meta-analytic review. Psychol. Assess. **14**(1), 16 (2002)

Bindler, A., Hjalmarsson, R.: Path dependency in jury decision making. J. Eur. Econ. Assoc. **17**(6), 1971–2017 (2019)

Du, Y., Clark, J.E.: New insights into statistical learning and chunk learning in implicit sequence acquisition. Psychon. Bull. Rev. **24**, 1225–1233 (2017)

Harris, C.R., Millman, K.J., van der Walt, S.J., et al.: Array programming with NumPy. Nature **585**, 357–362 (2020)

Hogarth, R.M., Einhorn, H.J.: Order effects in belief updating: the belief-adjustment model. Cogn. Psychol. **24**(1), 1–55 (1992)

Linn, R.L.: The standards for educational and psychological testing: guidance in test development. In: Handbook of Test Development, pp. 41–52. Routledge (2011)

Norris, J.R.: Markov Chains, No. 2. Cambridge university press, Cambridge (1998)

Nosofsky, R.M., Palmeri, T.J.: An exemplar-based random walk model of speeded classification. Psychol. Rev. **104**(2), 266 (1997)

Odic, D., Hock, H., Halberda, J: Hysteresis affects approximate number discrimination in young children. J. Exp. Psychol. Gen. **143**(1), 255 (2014)

Paulhus, D. L.: Measurement and control of response bias. In: Robinson, J.P., Shaver, P.R., Wrightsman, L.S. (eds.), Measures of Personality and Social Psychological Attitudes, pp. 17–59. Academic Press (1991)

Schuman, H., Presser, S.: Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. Sage, London (1996)

Schwarz, N., Sudman, S. (Eds.): Context Effects in Social and Psychological Research. Springer Science & Business Media, New York (2012). https://doi.org/10.1007/978-1-4612-2848-6

Sprent, P.: Data driven Statistical Methods. Routledge, London (2019)

Stewart, W.J.: Numerical methods for computing stationary distributions of finite irreducible Markov chains. In: Computational Probability, pp. 81–111. Springer, Boston, MA, US (2000)

Strang, G.: Introduction to Linear Algebra. Wellesley-Cambridge Press, Cambridge (2022)

Thiel, S.D., et al.: Hysteresis as an implicit prior in tactile spatial decision making. PLoS One **9**(2), e89802 (2014)

Integrating Rasch and Compositional Modeling for the Analysis of Social Survey Data

Antonio Calcagnì[1,2(✉)]

[1] University of Padova, Padua, Italy
antonio.calcagni@unipd.it
[2] GNCS, National Institute of Advanced Mathematics (INdAM), Rome, Italy

Abstract. Survey data are still ubiquitous in various fields of science, capturing human attitudes and opinions about relevant aspects of daily life. However, they often contain more information than typically conveyed. Understanding individual response processes can reveal instances of hesitancy and decision uncertainty, which shed light on the unobserved response mechanisms. We present a method to extract as much valuable insight as possible from survey responses using Item Response Theory tree and Compositional Data Analysis. Illustrated with a case study on reactions to the war in Ukraine, our approach provides an alternative framework for analyzing survey data.

Keywords: survey data · compositional data analysis · item response theory · dirichlet regression

1 Introduction

Rating data are prevalent across various scientific fields and allows for capturing human attitudes and opinion about social and individual phenomena. The measurement process, reliant on individuals, potentially holds more information than typically conveyed. Indeed, the responses provided in questionnaires only capture a portion of the response process, namely the terminal one. Tracing individual engagement in survey responses might eventually reveal varying hesitancy and decision uncertainty, potentially resulting in less or more extreme responses compared to those actually measured. As an example, consider a scenario where an individual is tasked with rating their satisfaction with their current work using a five-point scale. In this scenario, they undergo a process of gathering and integrating cognitive and affective information about their job satisfaction before settling on a response. However, this decision-making process may be fraught with uncertainty due to conflicting information, such as work-related problems which may negatively influence their response. It should be noted how this uncertainty does not originate from the nature of the question but rather from the intricate cognitive processes involved in formulating a response, which reflects the individual's epistemic state at the time of responding.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 87–92, 2025.
https://doi.org/10.1007/978-3-031-64346-0_15

In this work, we propose a method to extract valuable insight from survey responses and analyze them consistently. Our aim is to model respondent response processes using Item Response Theory tree [1] while analyzing the survey responses via Compositional Data Analysis (CoDa) [2]. We illustrate this approach with a case study on reactions to the war in Ukraine.[1].

2 A Rasch-IRTree Data Representation

The Rasch-tree model, a member of the IRTrees family [1], provides a straightforward statistical representation of rating responses using conditional binary trees [3]. Figure 1 depicts a standard IRTree example modeling a response scale with $M = 5$ anchors using N nodes.

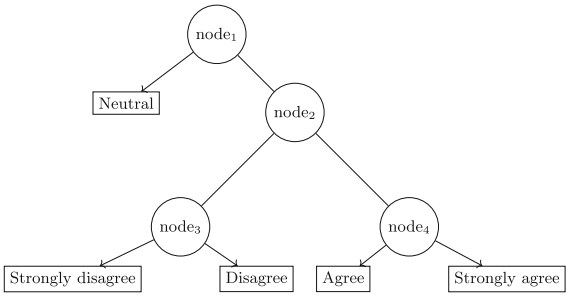


Fig. 1. Example of a five-point scale represented as a binary tree. The root node node_1 represents the starting point of the decision process, internal nodes represent switches between response states (e.g., node_2 switches between negative and positive response attitudes), endnotes represent the final responses.

Let $Y_{ij} \in \mathcal{C}$ be the random variable for the i-th respondent modeling one of the possible response from the response set $\mathcal{C}$ (with $|\mathcal{C}| = M$). Then, for a given survey question $j \in \{1, \dots, J\}$, the outcomes of Y_i can be rewritten using a dummy vector $\mathbf{y}_i^{\dagger} \in \{0, 1\}^M$, with $y_i^{\dagger} = 1$ only for the entry $Y_i = y$. Consequently,

$$\mathbb{P}(Y_i = y|\boldsymbol{\eta}_i; \boldsymbol{\alpha}) = \mathcal{M}ult(\mathbf{y}^{\dagger}; 1, \boldsymbol{\pi}_i^y)$$

The $M \times 1$ vector of probabilities $\boldsymbol{\pi}_i^y$ is mapped to the user-defined response tree model as follows:

$$\pi_{im}^y = \prod_{n=1}^{N} \left(\frac{\exp(\eta_{in} + \alpha_n) t_{mn}}{1 + \exp(\eta_{in} + \alpha_n)} \right)^{t_{mn}}, \qquad \boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$$

where t_{mn} is an entry of the Boolean mapping matrix $\mathbf{T}_{M \times N}$, which indicate how each response category (in rows) is associated to each node (in columns) of the

[1] All the materials like algorithms and datasets used throughout this contribution are available to download at https://github.com/antcalcagni/IRTree_CoDa.

tree, $\boldsymbol{\eta}_i \in \mathbb{R}^N$ expresses the rater's (latent) ability to answer the survey question, whereas $\boldsymbol{\alpha}_n \in \mathbb{R}^J$ expresses the easiness of choosing the n-node of the tree in the decision path. Note that both $\boldsymbol{\eta}_i$ and $\boldsymbol{\alpha}_n$ can vary among nodes. Finally, given a sample of responses $\mathbf{Y}_{n\times J}$, the parameters array $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\}$ can be estimated via maximum marginal likelihood [1]. Once the IRT parameters have been estimated, one can calculate the probabilistic model of the rater's multiverse $\tilde{\mathbf{y}}_{ij} \in [0,1]^M$. This model encapsulates the heterogeneity in the rater's response patterns and behaves as a model encompassing all potential responses the rater might achieve if they were to repeatedly answer a question within the same timeframe and under identical conditions. In this framework, decision uncertainty manifests through the estimated transition probabilities $\boldsymbol{\pi}_i^y$ wherein more certain responses necessitate smoother transitions among the tree's nodes.

Far from presenting findings as rigorously as Herzel [4], the IRTree procedure can be conceived as a form of *scale quantification*, which outputs a set of M probability masses for each respondent i and survey item j. At a final stage, researchers can either retain the probability masses as input for subsequent analyses or summarize them using scalars (e.g., by computing the expected value of the distribution). The first option paves the way for treating $\tilde{\mathbf{y}}_{ij}$ as compositional data.

3 Dirichlet Compositional Regression

Let $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{y}}_1, \ldots, \widetilde{\mathbf{y}}_i, \ldots, \widetilde{\mathbf{y}}_n)$ be a collection of independent random compositions with $\tilde{\mathbf{y}}_i \in \mathbb{S}^M$ and $\mathbb{S}^M = \{(y_1, \ldots, y_m, \ldots, y_M) \in \mathbb{R}_+^M | \sum_m y_m = 1\}$. Let $\mathbf{X}_{n\times J}$ be a matrix of observed variables (e.g., covariates), which can be used to predict the compositional response $\widetilde{\mathbf{Y}}$. The Dirichlet linear model with fixed dispersion is as follows:

$$\tilde{\mathbf{y}}_i \sim \mathcal{D}(y; \boldsymbol{\mu}_i\phi), \quad g(\boldsymbol{\mu}_i) = \mathbf{x}_i\boldsymbol{\beta}$$
$$\boldsymbol{\mu}_i = \left(\frac{1}{\sum_{m=1}^M \exp \mathbf{x}_i\beta_m}, \frac{\exp \mathbf{x}_i\boldsymbol{\beta}_2}{\sum_{m=1}^M \exp \mathbf{x}_i\beta_m}, \ldots, \frac{\exp \mathbf{x}_i\boldsymbol{\beta}_M}{\sum_{m=1}^M \exp \mathbf{x}_i\beta_m}\right)$$

where $\boldsymbol{\beta}_1 = \mathbf{0}_J$ acts as reference or base level whereas $g(.)$ is a well-defined link function (e.g., logit). Likewise for the Beta linear model also the Dirichlet regression is written using mean/dispersion parameterization (heteroscedasticity is explicitly considered in this case). The parameters estimation is performed via maximum likelihood methods, which also provides results for computing the expected Fisher information as well as the likelihood ratio test for models selection. We refer the reader to [5,6] for further details.

4 Case Study

We aimed to investigate the predictors of anxiety in watching the war in Ukraine in a sample of $n = 796$ participants from Canada, Germany, and Finland (68% female, mean age 24.4 years, 85% did not have relatives or friends involved

in the war).[2] The survey consisted of several questionnaires used to measure attitudes toward the war in Ukraine. For this research, we considered only a subset of items, retaining those measuring `anxiety` (six items) and `depression` (eight items), both assessed on 5-point scales, along with `gender`. The variable `anxiety` was quantified using the Rasch-IRTree procedure whereas `depression` was considered as total score.

To mitigate overfitting, the original dataset was partitioned into two equal-length subsets. One subset was utilized for estimating the IRTree parameters, while the other was reserved for the compositional data analysis. For the scale quantification, the decision tree depicted in Fig. 1 was used. IRTree parameters were estimated using a random-effect Binomial linear model [1]. Table 1 reports the estimates alongside their standard errors for each of the six item composing the outcome `anxiety`. Next, the estimates were used to compute the compositions for each rater and item, which were in turn averaged across the items to get the final compositional response $\widetilde{\mathbf{Y}}$. In this case, according to the Aitchinson's geometry, the perturbation average was used [2]. Figure 2 shows the ensuing compositional data for a selection of three respondents before averaging across items.

Table 1. Case study: Estimates ($\hat{\theta}$) and standard errors ($\sigma_{\hat{\theta}}$) for item parameters $\hat{\boldsymbol{\alpha}}$ alongside the estimated correlation matrix and standard deviations ($\hat{\sigma}_\eta$) for latent abilities $\hat{\boldsymbol{\eta}}$.

	node 1		node 2		node 3		node 4	
	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	$\hat{\theta}$	$\sigma_{\hat{\theta}}$
α_1	1.19	0.14	1.41	0.33	3.41	0.39	-3.35	0.38
α_2	1.13	0.14	-0.11	0.31	1.70	0.28	-3.28	0.40
α_3	1.60	0.15	1.37	0.32	2.96	0.35	-2.32	0.32
α_4	1.15	0.14	0.15	0.31	1.82	0.29	-2.60	0.36
α_5	1.62	0.15	0.70	0.31	1.81	0.30	-2.02	0.32
α_6	1.53	0.15	-1.16	0.31	0.95	0.24	-3.07	0.40

	η_1	η_2	η_3	η_4	$\hat{\sigma}_\eta$
η_1	1.00				1.12
η_2	-0.16	1.00			4.12
η_3	-0.38	0.79	1.00		1.76
η_4	0.46	0.59	0.20	1.00	2.32

Finally, to predict `anxiety` as a function of both `depression` and `gender`, we used the Dirichlet linear model with logit and log link functions for mean and precision components (the first response option was used as reference category). The mean component $\boldsymbol{\mu}_i$ included the main effects of `depression` and `gender` alongside their interaction `depression` $\times$`gender`. Figure 3 depicts the response variable as a function of both predictors whereas Table 2 shows the estimated parameters of the Dirichlet linear model whereas Fig. 4 plots the predicted curves against the compositional parts as a function of both continuous and categorical predictors. We observed that, as levels of depression increase, individuals are

[2] The dataset is publicly available at http://osf.io/whk48/. For further details about the survey, see [7].

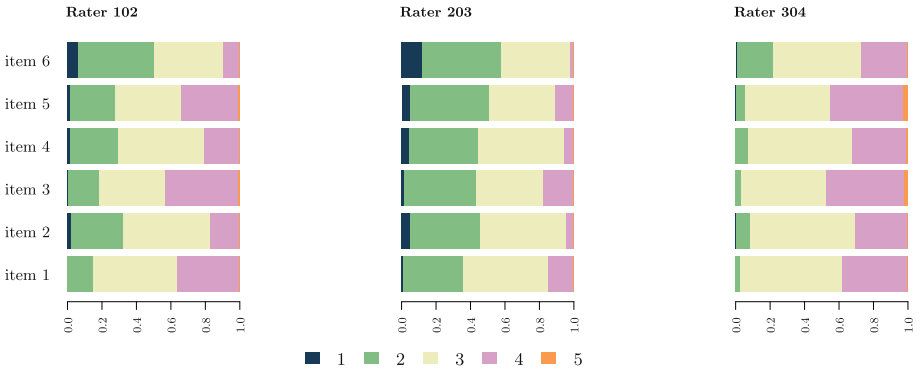


Fig. 2. Case study: Compositional responses for a selection of respondents. Note that response categories are represented with different colours.

more likely to self-report substantial anxiety compared to the baseline ($Y = 1$, *Not at all*). Specifically, the odds of experiencing extreme anxiety ($Y = 5$) increase by a factor of $\exp(\beta_{\text{depres}}) = 4.05$ compared to having no anxiety at all. Furthermore, we found an interaction between `depression` and `gender`. This interaction reveals that, compared to females with no anxiety, males are approximately 1.70 times more likely to experience moderate ($Y = 3$) or high ($Y = 4$) levels of anxiety when experiencing depression.

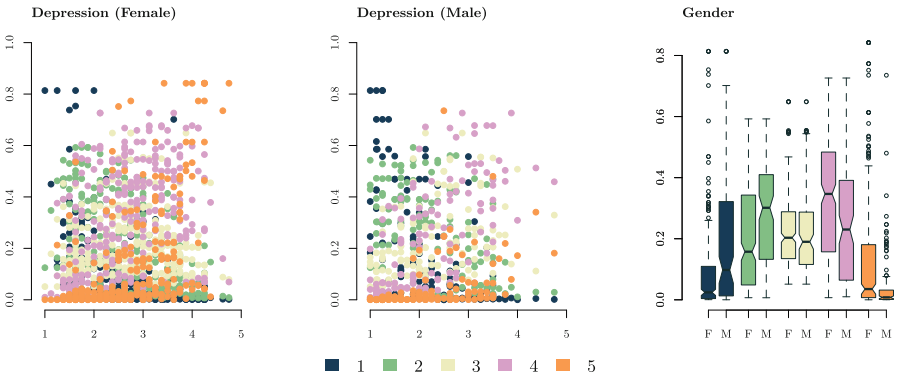


Fig. 3. Case study: Compositional responses as a function of `depression` and `gender`.

Table 2. Case study: Estimates ($\hat{\theta}$) and standard errors ($\sigma_{\hat{\theta}}$) for the Dirichlet linear model. Note that each response category is represented columnwise ($Y = 1$ is the reference category used to contrast the regression coefficients for $Y > 1$).

	$Y = 2$			$Y = 3$			$Y = 4$			$Y = 5$		
	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	z
β_0	0.57	0.25	2.28	–0.51	0.24	–2.08	–1.08	0.24	–4.54	–3.70	0.29	–12.87
β_{depres}	0.10	0.09	1.11	0.60	0.09	6.74	0.87	0.09	10.13	1.40	0.10	13.86
$\beta_{\text{gender:Male}}$	–1.04	0.37	–2.84	–1.36	0.37	–3.71	–1.64	0.36	–4.49	–0.27	0.43	–0.64
$\beta_{\text{depres}\times\text{gender}}$	0.41	0.15	2.69	0.51	0.15	3.41	0.57	0.15	3.85	–0.02	0.17	–0.11
ϕ	1.81	0.03	53.03									

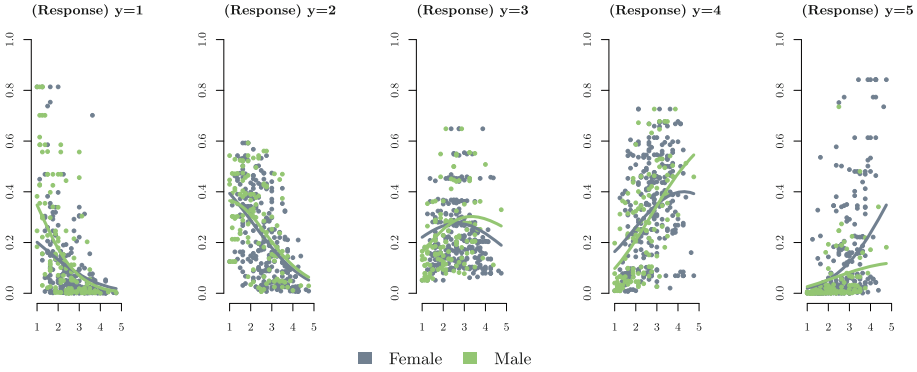


Fig. 4. Case study: Compositional predictions of the Dirichlet linear model.

Acknowledgements. The author acknowledges the financial support provided through the 2022 Italian MUR grant 2022APAFFN "The attentional curve of forgetting visual information in younger and older adults: experimental and computational insights", which supported this research.

References

1. Boeck, P.D., Partchev, I.: J. Stat. Soft. **48** (2012)
2. Filzmoser, P., Hron, K., Templ, M.: Springer, Cham (2018)
3. Jeon, M., Boeck, P.D.: Behav. Res. Methods **48**(3), 1070 (2015)
4. Herzel, A.: Metron **XXXII** (1974)
5. Gueorguieva, R., Rosenheck, R., Zelterman, D.: Computational statistics & data analysis **52**(12), 5344 (2008)
6. Maier, M.: Research Report Series, Vienna University of Economics and Business **125** (2014)
7. Greenglass, E., Begic, P., Buchwald, P., Karkkola, P., Hintsa, T.: Int. J. Psychol. (2023)

Artificial Neural Networks in Psychometrics Research

Monica Casella, Raffaella Esposito, Maria Luongo, Nicola Milano, Michela Ponticorvo, Roberta Simeoli, and Davide Marocco(✉)

Laboratorio NAC "Orazio Miglino" per lo studio della cognizione naturale ed artificiale, Dipartimento di Studi Umanistici, Università di Napoli "Federico II", Naples, Italy
{monica.casella,raffaella.esposito3,maria.luongo,nicola.milano, michela.ponticorvo,roberta.simeoli,davide.marocco}@unina.it

Abstract. The contribution of Artificial Neural Networks in psychometrics can help integrate the explanatory approach that often requires strong assumptions on data and the predictive one that, on the contrary, only needs mild assumptions on input data. Predictive techniques are able to identify data patterns and generate accurate predictions of output values starting from new sets of data, which is relevant in different psychometrics domains. Here we provide three examples of applications related to psychometric data analysis, language analysis and spatial cognition.

These studies show that integrating machine learning techniques into traditional psychometric data analysis allows to work on different input data, including not standard ones, to complement traditional procedures with new ones and to identify unexpected patterns.

Keywords: Artificial Neural Networks · Psychometrics · Data Analysis · LLMs · autism

1 Introduction

Traditional psychometric data analysis has historically relied on explanatory modeling techniques, aiming to examine hypothesized connections between variables and explore intercorrelations using methods like correlation or latent variable analysis. Using these techniques requires meeting strong assumptions.

In contrast, Artificial Neural Networks (ANNs) adopt a predictive approach, where researchers endeavor to uncover patterns in data and generate accurate predictions of output values based on input values of new observations. Unlike conventional methods, ANN techniques rely on fewer assumptions about input data.

Whereas explanatory modeling continues to be the prevailing approach, numerous studies have advocated for the integration of ANNs to enhance traditional psychometric methods. This integration aims to improve research efficiency, streamline model performance evaluation, and enhance interpretability. ANNs have found application across diverse realms of psychology, including psychometrics and the related issues of psychometric test validation. The subsequent section offers three examples of these applications.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 93–98, 2025.
https://doi.org/10.1007/978-3-031-64346-0_16

1.1 Artificial Neural Networks for Psychometric Data Analysis

Traditionally, psychometric data analysis has heavily relied on explanatory techniques for tasks such as item selection, scale validation, and missing data imputation. However, recent studies have underscored the potential benefits of integrating Artificial Neural Networks (ANNs) into this domain. ANNs offer significant advantages due to their flexibility and the diverse architectures and configurations they can assume, making them adaptable to a wide range of scenarios.

In the realm of item selection and scale validation, ANNs represent a promising approach for maximizing the predictive capability of psychometric scales and models and for overcoming the assumptions of statistical models commonly used in this context.

A study conducted by Dolce et al. [1] exemplified the integration of both explanatory and predictive perspectives in item selection. This study proposed a methodology that combines exploratory data analysis for investigating dimensional structures with ANNs for predicting psychopathological diagnoses, effectively treating it as a classification problem. This approach not only provides theoretical insights into the characteristics of selected items but also enhances predictive accuracy regarding external criteria, that is the psychopathological diagnoses.

Similarly, Casella et al. [2] applied ANNs in developing shortened versions of psychological tests. While creating short forms traditionally involves laborious processes and limited exploration of alternatives, ANNs offer automation and optimization, selecting short forms that best reconstruct responses from the original long test while preserving its dimensionality.

Scale validation often involves studying the internal structure of tests, commonly done through factor analysis and related techniques. However, these methods impose linear dimensionality reductions under assumptions that may not always hold in psychological data. Predictive techniques like ANNs offer a complementary approach, providing the possibility to explore latent spaces in a nonlinear fashion and potentially improving the understanding of item-factor relationships, as demonstrated in studies by Casella et al. [3, 4] Esposito et al. [5], and Milano et al. [6].

ANNs also show promise in handling missing data, a common challenge in psychometric analysis, particularly with Likert-scale responses. Conventional imputation techniques, such as discriminant analysis or logistic regression, may introduce biases, especially with asymmetric items. Caution is advised with methods like listwise deletion or mean imputation due to their unrealistic assumptions in survey and rating scale data. ANNs have exhibited considerable potential in addressing missing data across various mechanisms when compared to alternative imputation techniques [7, 8].

In summary, the integration of Artificial Neural Networks into psychometric data analysis represents a flourishing field with the potential to enhance both the validity and the predictive power of psychological models.

1.2 Semantic Similarity Between Items Embeddings

Large language models (LLMs) [9] have become rapidly the state-of-the-art method to understand and generate human comprehensive text. Commercial models like ChatGPT and Google Gemini are nowadays commonly used by millions of people in an enormous

variety of tasks. In the context of psychological science LLMs have generated interest for their capacity to understand and contextualize human language and to generalize to previously unseen texts in zero-shot [10–12].

Here we describe how to exploit LLMs to retrieve semantic similarity from test items, and if this similarity predicts in some way the factorial structure that we hypothesize, and that then we find in participants' responses.

LLMs, in their internal representation, map the input text into vectors. These vectors are learnt during the training phase and are called embeddings. This way a sentence, or just a single word, becomes a floating-point vector of fixed size. There is evidence that the embedding space has metric property and that similar concepts, i.e. colors or synonymous, are mapped closer in the embedding space than different ones [13]. Exploiting this property, embeddings can be used to measure semantic similarity in different domains [13, 14]. Here we propose to use traditional validated psychological test embeddings to find semantic similarity between items and check if the test follows the expected factorial structure looking only at the items, and not at the responses.

To achieve this goal, we use Bidirectional Encoder Representations from Transformers (BERT), [15] a pre-trained language model released from Google and completely open source. BERT is trained on a dataset made of 5 billion sentences from Wikipedia and Google Books corpus. Its goal is to predict words from masked sentences. Since its release in 2018, several modifications of the basic BERT model have been proposed. We used the one with the higher score on the LLMs standard benchmarks table [16]: roBERTa.

As a preliminary approach we use the Big5 test [17], we map the items in the embedding space using BERT and obtain for each of the 50 items its embedded representation in the vector space, composed of 1024 dimensions. To measure how much embeddings representing different items are closer in the embedding space we use the cosine similarity.

Cosine similarity is the cosine of the angle between the vectors; that is, it is the dot product of the vectors divided by the product of their lengths. It follows that the cosine similarity does not depend on the vectors' magnitudes, but only on their angle.

$$cosinesimilarity(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\| \|\boldsymbol{B}\|} \tag{1}$$

Here we report the cosine similarity equation, where A and B are two items' embeddings. Thus, we obtain a similarity matrix that can be used for different targets. We can cluster the matrix to see if our embeddings follow the expected underlying factorial structure, or we can directly apply a factorial analysis to the cosine similarity matrix, that when the vectors are centre, have zero mean, is the same as the Pearson correlation coefficient. In conclusion, looking at the items only and exploiting LLMs property, we can infer if our test structure adheres to the expected factorial structure. Furthermore, we can change items that do not correlate with the others, or we can delete items representing the same construct to shorten the test.

1.3 Digital Technology and AI to Identify Children with Autism Spectrum Disorder

Recent studies have highlighted the significant role of motor behavior in understanding and diagnosing Autism Spectrum Disorder (ASD). This has led to the integration of digital technology with artificial intelligence, fostering the development of interactive assessments on smart tablets [18]. Such an approach provides a child-friendly and ecologically valid method for assessing motor skills, simultaneously generating a rich dataset for in-depth analysis. In particular, leveraging the data gathered from tablet-based activities, researchers have employed artificial neural networks to analyze motor patterns, demonstrating remarkable efficacy in differentiating between children with ASD and their typically developing peers [19, 20].

Specifically, Simeoli et al. [19] developed a software tool that through a smart tablet device and touch screen sensor technologies recorded motion patterns. An Artificial Neural Network was used to classify the movement trajectories of autistic children and typically developing children. Results revealed a 93% classification accuracy, demonstrating that autism can be computationally identified. The analysis of the features that most affect the prediction reveals and describes the differences between the groups, confirming that motor abnormalities are a core feature of autism.

Milano et al. [20] conducted further research to closely observe the behavior of the features involved in classification. They applied a variational autoencoder, a particular type of Artificial Neural Network to observe the latent distribution description of motion features. Their results revealed that the motion features of children with autism consistently differ from those of children with typical development, suggesting that it could be possible to identify potential motion hallmarks typical for autism and support clinicians in their diagnostic process, using these innovative systems.

Building on these insights, Luongo et al. [21] conducted a study using a smart tablet with game-based software to examine the motor patterns of children through a drag-and-drop task. This study compared the movement trajectories between children diagnosed with ASD and typically developing children. By focusing on sequential and raw data, this research unveils a fresh perspective in identifying distinctive markers of autism, thereby making a significant contribution to the field of ASD research. Indeed, this methodology not only provides a detailed quantitative framework for assessing and interpreting motor patterns in children with autism but also sets the stage for more accurate and earlier diagnoses.

2 Conclusions

This paper examines the utilization of ANNs in the domain of psychometrics. While conventional psychometric techniques rely on explanatory modeling, ANNs offer an alternative approach wherein researchers seek to discern patterns within data and produce precise forecasts of output values based on input values from new observations.

We have explored three different approaches to the psychometric domain, including data analysis, language analysis and autism classification. These examples show that ANNs have the potential to advance in the integration of the explicatory and predictive modelling approaches.

The ANNs capacity to furnish precise predictions based on new data inputs is especially vital in psychology, where accurate prediction of outcomes is imperative for informed decision-making and effective intervention strategies. By leveraging ANNs to discern patterns in data, researchers can formulate more accurate models of human behavior and cognition, thereby aiding the development of novel interventions. Moreover, they are adept at handling extensive datasets and intricate relationships among variables. This capability proves particularly advantageous in fields such as psychology, where datasets often exhibit complexity, challenging traditional statistical methodologies.

ANNs, in particular, exhibit significant promise across a spectrum of psychological applications, ranging from the creation and validation of psychometric tests to the identification of neurological disorders like autism. By harnessing these potent algorithms, researchers can glean novel insights into the underlying mechanisms of measurement and diagnosis, thus paving the way for more efficacious interventions.

To sum up, integrating ANNs into conventional psychometric data analysis and overall methodology holds the potential to furnish novel avenues for data analysis and interpretation, giving new insights into psychometrics.

References

1. Dolce, P., Marocco, D., Maldonato, M.N., Sperandeo, R.: Toward a machine learning predictive-oriented approach to complement explanatory modeling. An application for evaluating psychopathological traits based on affective neurosciences and phenomenology. Front. Psychol. **11** (2020)
2. Casella, M., Dolce, P., Ponticorvo, M., Milano, N., Marocco, D.: Artificial neural networks for short-form development of psychometric tests: a study on synthetic populations using autoencoders. Educ. Psychol. Measur. **84**, 62–90 (2024). https://doi.org/10.1177/00131644231164363
3. Casella, M., Dolce, P., Ponticorvo, M., Marocco D.: Autoencoders as an alternative approach to principal component analysis for dimensionality reduction. An application on simulated data from psychometric models. In: PSYCHOBIT (2021)
4. Casella, M., Dolce, P., Ponticorvo, M., Marocco, D.: From principal component analysis to autoencoders: a comparison on simulated data from psychometric models. In: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 377–381. IEEE, Rome, Italy (2022). https://doi.org/10.1109/MetroXRAINE54828.2022.9967686
5. Esposito, R., Casella, M., Milano, N., Marocco, D.: Autoencoders as a tool to detect nonlinear relationships in latent variables models. In: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 1012–1016 (2023). https://doi.org/10.1109/MetroXRAINE58569.2023.10405761
6. Milano, N., Casella, M., Esposito, R., Marocco, D.: Exploring the Potential of Variational Autoencoders for Modeling Nonlinear Relationships in Psychological Data. (submitted)
7. Sun, Y., Li, J., Xu, Y., Zhang, T., Wang, X.: Deep learning versus conventional methods for missing data imputation: a review and comparative study. Expert Syst. Appl. (2023)
8. Collier, Z.K., Kong, M., Soyoye, O., Chawla, K., Aviles, A.M., Payne, Y.: Deep learning imputation for asymmetric and incomplete Likert-type items. J. Educ. Behav. Stat. **49**, 241–267 (2024). https://doi.org/10.3102/10769986231176014

9. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
10. Binz, M., Schulz, E.: Turning large language models into cognitive models. arXiv preprint arXiv:2306.03917. (2023)
11. Buschoff, L.M.S., Akata, E., Bethge, M., Schulz, E.: Have we built machines that think like people?. arXiv preprint arXiv:2311.16093. (2023)
12. Chuang, Y.S., et al.: Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv preprint arXiv:2311.09618. (2023)
13. Yan, F., Fan, Q., Lu, M.: Improving semantic similarity retrieval with word embeddings. Concurr. Comput.: Pract. Exp. **30**(23), e4489 (2018)
14. Colla, D., Mensa, E., Radicioni, D.P.: Novel metrics for computing semantic similarity with sense embeddings. Knowl.-Based Syst. **206**, 106346 (2020)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018)
16. https://www.sbert.net/docs/pretrained_models.html
17. McCrae, R.R., Costa, P.T.: Updating Norman's adequacy taxonomy: intelligence and personality dimensions in natural language and in questionnaires. J. Pers. Soc. Psychol. **49**(3), 710 (1985)
18. Luongo, M., Simeoli, R., Marocco, D., Ponticorvo, M.: The design of a game-based software for children with autism spectrum disorder. In: 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 318–322. IEEE, October 2022
19. Simeoli, R., Milano, N., Rega, A., Marocco, D.: Using technology to identify children with autism through motor abnormalities. Front. Psychol. **12**, 635696 (2021)
20. Milano, N., Simeoli, R., Rega, A., Marocco, D.: A deep learning latent variable model to identify children with autism through motor abnormalities. Front. Psychol. **14**, 1194760 (2023)
21. Luongo, M., Simeoli, R., Marocco, D., Ponticorvo, M.: Exploring motor patterns in autism spectrum disorder using raw data and artificial intelligence: a pilot study. In: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), pp. 1006–1011. IEEE, October 2023

Sparse Partial Membership Models with Applications in Food Science

Alessandro Casa[1(✉)], Thomas Brendan Murphy[2], and Michael Fop[2]

[1] Department of Economics, University of Bergamo, Bergamo, Italy
alessandro.casa@unibg.it
[2] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
{brendan.murphy,michael.fop}@ucd.ie

Abstract. With the growth of consumers' awareness towards food quality and sustainability, we are witnessing an increased demand for tools capable of detecting food adulteration. In this framework, vibrational spectroscopy allows for the rapid collection of vast amount of highly informative data to be used in food authenticity studies. This paper introduces a sparse partial membership model for food adulterant identification using spectrometry data, which are high-dimensional and characterized by complex relations among the variables. The proposal not only enables the identification of adulterated samples but also detects the percentage of adulterant while determining which spectral regions are more impacted by it. This could lead to richer chemical insights and to the development of faster portable instrument to collect data to be subsequentially used in food authenticity studies.

Keywords: Partial membership · penalized likelihood · food authentication · chemometrics

1 Introduction

In the evolving food production landscape, there is a growing awareness about food quality, traceability and sustainability among consumers, retailers and food processors. This is highlighting the importance of food integrity and the need for tools capable of detecting food adulteration. Adulteration is defined as the act of removing or replacing food original components with cheaper alternatives, and it might have both economic and health concerning implications [4].

Historically, food authenticity studies implemented to verify whether food products match their purported identity, have employed time-consuming and expensive laboratory processes. More recently, vibrational spectroscopy techniques have become an efficient and relatively cheap alternative to collect data to be used for food authenticity purposes. From a statistical perspective, these data pose some challenges mainly related to their high-dimensionality and to the intricate correlation structures among the observed variables, referred to as wavelengths.

In this context, individual-level mixture models [1], and in particular partial membership models [6, PMM], emerge as a promising approach for food authenticity studies. Differently from standard mixture-based clustering approaches, these models allow

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 99–104, 2025.
https://doi.org/10.1007/978-3-031-64346-0_17

each unit to belong simultaneously to different components with different degrees of memberships. Unfortunately, PMM tendendency to be over-parameterized in high-dimensional settings jeopardizes their usefulness when analyzing spectroscopy data. For this reason, in this work we propose a sparse PMM, which is based on the adoption of appropriate penalty functions. The resulting model introduces a refined authentication tool capable to identify not only if a sample has been adulterated or not but also the percentage of adulteration. At the same time, the induced sparsity allows for a parsimonious description of the relations among the wavelengths and the identification of spectral regions most influenced by the adulterant. This can lead to relevant insights from an applied viewpoint and could constitute a starting point for further chemical analyses. Moreover, it might enhance the possibility to build new portable spectrometers that can collect data in a cheaper and faster way.

The rest of the paper is structured as follows. In Sect. 2, we introduce our proposed method and outline the devised estimation procedure. In Sect. 3, the method is tested on real spectroscopy data. Lastly, Sect. 4 concludes the paper with some considerations and avenues for future research.

2 Proposed Methodology

Let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, with $\mathbf{y}_i \in \mathbb{R}^p$, be the sample of the observed data. According to the PMM formulation, and adopting the standard parameterization for the multivariate Gaussian distribution, $\mathbf{y}_i$ is distributed as

$$(\mathbf{y}_i|\mathbf{g}_i, \Theta) \sim N_p \left(\left(\sum_{k=1}^{K} g_{ik} \Sigma_k^{-1} \right)^{-1} \left(\sum_{k=1}^{K} g_{ik} \Sigma_k^{-1} \mu_k \right), \left(\sum_{k=1}^{K} g_{ik} \Sigma_k^{-1} \right)^{-1} \right), \quad (1)$$

where $\Theta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ denotes the collection of mixture component means and covariance matrices, while $\mathbf{g}_i = (g_{i1}, \ldots, g_{iK})$ is the partial membership vector for the i-th observation, with $g_{ik} \in [0, 1]$ for $k = 1, \ldots, K$ and constrained to $\sum_k g_{ik} = 1$.

To frame model (1) in the considered applied scenario, we assume that $K = 2$, with the two components corresponding to *pure food* and *pure adulterant*. Moreover, we assume that $\mu_1 = \mu^P = (\mu_1, \ldots, \mu_p)$ and $\mu_2 = \mu^A = (\mu_1 + \delta_1, \ldots, \mu_p + \delta_p)$ with μ^P and μ^A the mean vectors for the pure food and adulterant, respectively. The *mean-shifts vector* $\delta \in \mathbb{R}^p$ encodes differences between μ^P and μ^A; while additivity may look simplistic, it finds chemical confirmation in the so-called *Beer-Lambert law* [2] and is rather flexible in practice. Lastly, we assume that $\Sigma_1 = \Sigma_2 = \Sigma$. Consequently, model (1) can be rewritten as follows

$$(\mathbf{y}_i|\mathbf{g}^A, \Theta^*) \sim N_p(\mu^P + \delta g_i^A, \Sigma), \quad (2)$$

where $\Theta^* = \{\mu^P, \delta, \Sigma\}$ and $\mathbf{g}^A = (g_1^A, \ldots, g_n^A)$, with $g_i^A = g_{i2}$ representing the percentage of adulterant for the i-th sample.

With the aim of accommodating the peculiar characteristics of spectroscopy data, we assume that the parameters are sparse, and introduce appropriate penalty functions

in the estimation procedure to induce sparsity. Estimates are obtained by maximizing the following penalized log-likelihood

$$\ell_p(\mathbf{g}^A, \Theta^*; \mathbf{Y}) = \sum_{i=1}^{n} \log \phi(\mathbf{y}_i; \mu^P + \delta g_i^A, \Omega) - p_\lambda(\mathbf{g}^A, \delta, \Omega)\,, \tag{3}$$

where $\phi(\cdot; \mu^P + \delta g_i^A, \Omega)$ is the density of a multivariate Gaussian distribution with mean $\mu^P + \delta g_i^A$ and precision matrix $\Omega = \Sigma^{-1}$. The second term in (3) is a penalty term on the model parameters. Given the aim of detecting levels of the adulterant, selecting informative wavelengths, and account for high-dimensionality, we consider

$$p_\lambda(\mathbf{g}^A, \delta, \Omega) = \lambda_g ||\mathbf{g}^A||_1 + \lambda_\delta ||\mathbf{D}\delta||_1 + \lambda_\Omega ||\Omega||_1\,, \tag{4}$$

with $\lambda = (\lambda_g, \lambda_\delta, \lambda_\Omega)$ controlling the penalization strength and $||\cdot||_1$ the L_1-norm.

The specific choices adopted for the penalties in (4) are motivated by the application. In fact, the first term allows to shrink exactly to zero some g_i^A, thus providing a first discrimination between adulterated and non-adulterated samples. The second term in (4) impose a generalized lasso penalty [9], where different specifications of the matrix $\mathbf{D}$ encompass different well-known sparsity inducing penalties employed for variable selection. Here we specify $\mathbf{D}$ to place a sparse fused lasso penalty on δ. This choice not only allows to set to zero some of the elements in δ but it shrinks to zero the differences $|\delta_{j+1} - \delta_j|$, for $j = 1, \ldots, (p-1)$; this allows for the automatic detection of variables that are not impacted by the adulterant, while simultaneously accounting for high correlation between adjacent wavelengths, which tend to behave similarly. Lastly, the third term in (4) is a graphical lasso penalty [5] on Ω; this allows to reduce the number of free parameters to be estimated and to obtain a convenient and visual interpretation in terms of conditional independence among wavelengths, thanks to the connection with *Gaussian graphical models*.

2.1 Model Estimation and Model Selection

Hereafter, we outline the procedure to estimate Θ^* and $\mathbf{g}^A$. Note that, coherently with food authenticity applications where we know the food being analyzed, μ^P is considered as known. Estimates for the remaining parameters, $\mathbf{g}^A, \delta$ and Ω, are obtained by maximizing (3) iteratively alternating partial optimization steps.

In particular, the update for the mean-shift vector δ, considering Ω and $\mathbf{g}^A$ fixed, is carried out by means of the alternating direction method of multipliers [3, ADMM] which, by resorting to a variable splitting scheme, often leads to faster optimization in penalized settings. On the other hand, the update of the precision matrix Ω is obtained by employing the coordinate descent graphical lasso algorithm as proposed in [5]. Lastly, $\mathbf{g}^A$ is updated by means of a soft-thresholding operator with the additional constraint of forcing the estimates to lie in the interval $[0, 0.5]$, where 0.5 is assumed to be the maximum possible level of adulteration. The three partial optimization steps are iterated until a convergence criterion on the relative improvement of (3) is met.

The devised estimation procedure relies on the selection of reasonable values for the hyperparameter λ. In this work, we resort to model selection tools and we select

$\lambda_g, \lambda_\delta$ and λ_Ω by means of a modification of the Bayesian Information Criterion (BIC) conceived for penalized maximum likelihood. Operationally, two approaches are being explored. The first consists in relying on grid search over reasonable values for $\lambda_g, \lambda_\delta$ and λ_Ω and to select the best combination of hyperparameters, i.e. the one leading to the model with the highest BIC. The second solution consists in a sequential greedy scheme where elements of λ are optimized one at a time, keeping the others fixed to the values chosen at the previous iteration. The search is ended when changes in λ do not lead to an improvement of the BIC value. This turns out to be particularly suitable when an exhaustive grid search is computationally too demanding.

3 Real Data Application

Our proposal is tested on mid-infrared (MIR) spectroscopic measurements of honey samples [7]. In the original data, contamination with five different sugar syrups have been implemented; however, here we restrict our attention to beet sucrose adulterated samples. Therefore, in the final dataset we have $n = 410$ spectra, $n_H = 290$ from pure honey and $n_B = 120$ adulterated with different percentages (10%, 20%, 30%) of beet sucrose. Each spectrum consists of absorbance values measured at $p = 285$ wavelengths in the range 3700 nm–13600 nm. However, to reduce the computational burden, in the following analyses a variable aggregation step has been performed reducing the number of wavelengths to 95; as pointed out by [8], aggregation of adjacent and strongly correlated wavelengths implies almost negligible information losses. Moreover, a certain degree of supervision have been employed during the estimation procedure; specifically, we assumed to know g_i^A for 40 randomly chosen spectra among the n observed ones. Lastly, the optimal hyperparameter vector λ has been selected by resorting to the first strategy outlined in Sect. 2.1.

For the selected model, four elements of the estimate of the shift vector $\hat{\delta}$ are shrunk to zero, indicating that four of the aggregated wavelengths are not impacted by the adulterant. The procedure is also able to retrieve the presence of some regions where the mean-shifts are constant. Furthermore, to compare $\hat{\delta}$ and δ, the mean-shifts estimated assuming the adulteration levels as known, we resort to the *root mean squared error*

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{j=1}^{p}(\hat{\delta}_j - \delta_j)^2}\,.$$

In this case, RMSE $= 0.006$, thus indicating an adequate estimation of δ.

In Fig. 1, the estimated adulteration levels $\hat{g}_i^A$, for $i = 1, \ldots, n$, are displayed alongside with the true levels of adulteratio: none (0%), 10%, 20% and 30%. Our proposal provides excellent results in terms of recovery of the honey contamination, with a slight over-estimation for the adulterated samples with the highest degree of adulteration. The penalization strategy visibly allows to shrink some of the $\hat{g}_i^A$ exactly to zero, with 71.7% of the pure honey samples being correctly recognized as non-adulterated. As a further confirmation of the quality of the results, we compare $\hat{g}_i^A$ with the true g_i^A by means of the *mean absolute error*

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{g}_i^A - g_i^A|\,.$$

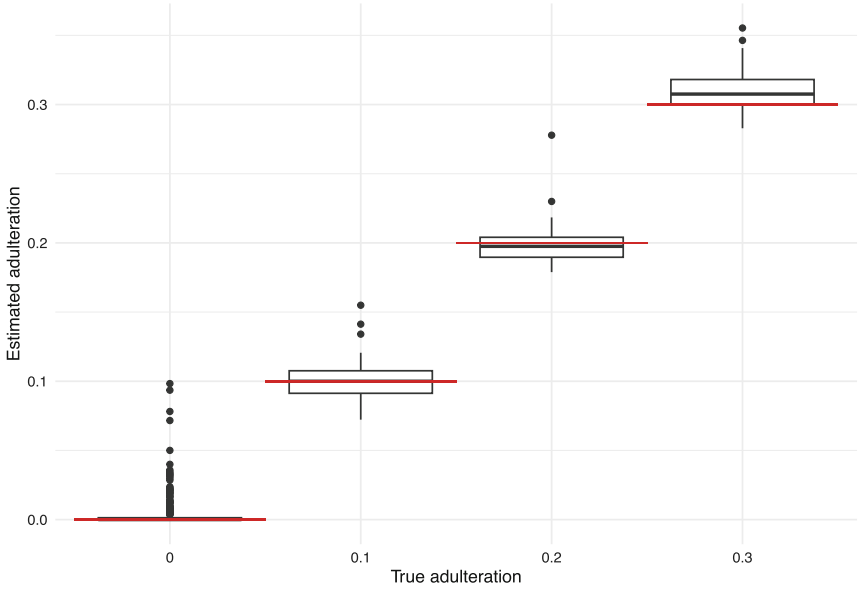


Fig. 1. Boxplot of the estimated adulteration levels $\hat{g}_i^A$ for each true adulteration level, highlighted with the red horizontal segments.

We obtain MAE $= 0.006$, highlighting that on average we make an estimation error of less than 0.6% in assessing the level of adulteration.

Lastly, the exploration of the estimated precision matrix $\hat{\boldsymbol{\Omega}}$ has shown that the graphical lasso allowed to greatly reduce the number of free parameters. The resulting matrix, being almost diagonal, confirms previous knowledge about the sparse structure of the partial correlations among the wavelengths when analysing spectroscopy data.

4 Conclusion and Discussion

We presented a sparse partial membership model that allows the analysis of high-dimensional spectroscopy data for food authentication purposes. The method moves a step forward with respect to standard classification-oriented approaches, which usually allow only the discrimination between adulterated and non-adulterated samples, without providing indication of the degree of the adulteration. In fact, by relying on partial membership models, our proposal is able to detect the sample-specific percentage of adulteration, thus providing richer information. Moreover, by means of appropriate sparsity inducing penalties, the method is capable of automatically detecting which spectral regions are more influenced by the adulterant, potentially shedding light on the underlying chemical processes. The method has been tested on mid-infrared spectroscopic data from pure and beet-sucrose adulterated Irish honey samples, showing results of good quality and satisfactory detection of the different degrees of adulteration.

This work opens up directions for future research. Firstly, the penalized procedure could be extended to deal with sparse covariance matrices rather than with precision

matrices; in fact, connections with *Gaussian covariance graph models* would retain a convenient interpretation in terms of marginal independencies between wavelengths. Alternatively, it would be interesting to consider eigen-decompositions of the covariance matrices, as these decompositions are widely employed in the model-based clustering literature. Lastly, the possibility to frame our proposal in a fully Bayesian framework is currently under exploration.

References

1. Airoldi, E.M., Blei, D., Erosheva, E.A., Fienberg, S.E.: Handbook of Mixed Membership Models and Their Applications. CRC Press, Boca Raton (2014)
2. Beer. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. Annalen der Physik, **162**(5), 78–88 (1852)
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**(1), 1–122 (2011)
4. Dimitrakopoulou, M.E., Vantarakis, A.: Does traceability lead to food authentication? A systematic review from a European perspective. Food Rev. Int. **39**(1), 537–559 (2023)
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
6. Heller, K.A., Williamson, S., Ghahramani, Z.: Statistical models for partial membership. In: Proceedings of the 25th International Conference on Machine Learning, pp. 392–399 (2008)
7. Kelly, J.D., Petisco, C., Downey, G.: Application of Fourier transform midinfrared spectroscopy to the discrimination between Irish artisanal honey and such honey adulterated with various sugar syrups. J. Agric. Food Chem. **54**(17), 6166–6171 (2006)
8. Murphy, T.B., Dean, N., Raftery, A.E.: Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Ann. Appl. Stat. **4**(1), 396–421 (2010)
9. Tibshirani, R.J., Taylor, J.: The solution path of the generalized lasso. Ann. Stat. **39**(3), 1335–1371 (2011)

Study of Clustering Algorithms for Mixed-Type Data in Presence of Errors and Correlation

Valentina Veronesi(✉) and Marianthi Markatou

Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo, Getzville, USA
vverones@buffalo.edu

Abstract. Clustering mixed-type datasets containing continuous, ordinal, nominal, and binary variables poses significant challenges, especially in the presence of measurement error (ME) and misclassification (Mi). This study examines the impact of ME and Mi on the clustering results of algorithms designed specifically for mixed-type data in the case of data with correlation. The clustering algorithms under evaluation are k-prototypes, Modha-Spangler, KAMILA, HyDaP, and PDQ. By highlighting the influence of data inaccuracies on these methods, our research aims to underscore the importance of choosing robust clustering techniques for analyzing complex, real-world information. The result of this study not only sheds light on the resilience of each algorithm to ME and Mi, but also guides practitioners in selecting the most appropriate clustering methods for their specific data challenges.

Keywords: Clustering mixed-type data · measurement error and misclassification · correlation

1 Introduction

Cluster analysis techniques, unsupervised and exploratory in nature, are concerned with exploring data sets to assess whether or not they can be meaningfully summarized in terms of a relatively small number of groups of objects that are similar to each other and different from individuals in other clusters [1]. Cluster analysis thus looks for a latent categorical variable, the group membership, that embeds the possible underlying structure.

Mixed-type data are data that comprise realizations of continuous, ordinal, nominal, and binary variables and are frequently met in practice. For example, clinical data sets may contain variables such as blood pressure (interval scale), death status (binary), blood type (categorical), and pain on a Likert scale (ordinal).

Clustering mixed-type data presents its own challenges, discussed in [2]. Collected data may be incorrect, as well. Measurement error occurs whenever it is not possible to observe accurately one or more variables that are considered

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 105–110, 2025.
https://doi.org/10.1007/978-3-031-64346-0_18

in a study [3]. The term *measurement error* (ME) is generally used for interval scale variables, while in the case of categorical variables it is referred to as *misclassification* (Mi).

Clustering often assumes variable independence, explicitly in model-based methods and implicitly through distance metric choices in distance-based approaches. In the local independence assumption, variables are independent given cluster membership. Similarly, the global independence assumption considers independence between informative and non-informative variables. Accounting for correlation when dealing with mixed-type variables is challenging [4] and literature is scarce in the case of clustering mixed-type data in the presence of correlated variables (e.g., [5]).

In this paper, we investigate the impact of ME and Mi on clustering mixed-type correlated data. Section 2 succinctly describes the algorithms we study, Sect. 3 discusses the models we use for ME and Mi, while Sects. 4 and 5 describe our simulation study and results. Section 6 presents the conclusions of our study.

2 Selected Algorithms

The studied algorithms were selected because of their accessibility, i.e., availability of an associated software, innovativeness, and ability to represent a diverse range of clustering methods suitable for mixed-type data. Table 1 summarizes the main characteristics of the algorithms.

3 Measurement Error and Misclassification Models

Let N represent the total number of observations and P the total number of variables in the data matrix $\mathbf{X}$ of dimensions $N \times P$, aiming for K clusters. Here, P is divided into H interval scale variables and M categorical variables, such that $P = H + M$. For each observation $\mathbf{X}_i = (\mathbf{V}_i, \mathbf{W}_i)$, $i = 1, \ldots, N$, $\mathbf{V}_i$ denotes interval data and $\mathbf{W}_i$ categorical data, with L_m the number of levels for the m-th categorical variable. When ordinal variables are included, $P = H + M + O$ and $\mathbf{X}_i = (\mathbf{V}_i, \mathbf{W}_i, \mathbf{Y}_i)$, incorporating ordinal data $\mathbf{Y}_i$.

Suppose a variable X is measured with error, with the true value of X being unobserved. We denote the error-prone variable by $\tilde{X}$. The relationship between the unobserved X and the observed $\tilde{X}$ is specified by the classical error model for continuous variables, given as $\tilde{V} = V + U$, with a random error U such that $E[U] = 0$ and $U \perp X$.

Consider a categorical variable W, with $L \geq 2$ categories with probability π_l, $l = 1, \ldots, L$, that is $P(W = w_l) = \pi_l$, $l = 1, \ldots, L$ with the constrain $\sum_{l=1}^{L} \pi_l = 1$. Let $\boldsymbol{\pi}^t = (\pi_1, \ldots, \pi_L)$, and note that $\pi_L = 1 - \sum_{l=1}^{L-1} \pi_l$. The observed $\tilde{W}$ could have a different number of categories, but in the following, there is the assumption that $\tilde{W}$ has the same number of categories as W, with misclassification probabilities $P(\tilde{W} = \tilde{w}_l | W = w_l) = \theta_{\tilde{w}_l|w_l}$. As the π_l, the $\theta_{\tilde{w}_l|w_l}$ are such that $\sum_{l=1}^{L} \theta_{\tilde{w}_l|w_l} = 1$. The entries of the matrix $\boldsymbol{\Theta}$, of elements $\theta_{\tilde{w}|w}$,

Table 1. Short description of the studied algorithms. Only the *principal* stopping rule for each algorithm is detailed; algorithms halt at a maximum iteration count if this rule is not met. PAM stands for Partition Around Medoids [12]. *Software* refers to the R programming language.

Algorithm	Characteristics (Method; Initialization; Objective function; Stopping rule; Software)
k-prototypes [6]	Distance-based; Random clusters' centers selection from data; Minimization of squared Euclidean distance for continuous variables and a weighted mismatch distance for categorical variables; Cluster assignment stabilization or objective function under threshold; `kproto` in `clustMixType` package [7].
M-S [8]	Distance-based; Random clusters' centers selection from data; Minimization of weighted sum of squared Euclidean distance for standardized continuous variables and cosine distance for standardized, dummy-coded categorical variables, with automatically selected weights; Difference of consecutive objective functions under threshold; `gmsClust` in `kamila` package [9].
KAMILA [10]	Semi-parametric; Random selection of centroids for interval variables and Dirichlet distribution for categorical parameters; Minimization of the product of categorical negative log-likelihood and the continuous variables' within-to-between cluster distance ratio; Clster assignment stabilization; `kamila` in `kamila` package [9].
HyDaP [11]	Distance-based; Greedy; PAM objective function with standardized Gower's distance; Cluster assignments stabilization; https://github.com/gmailw1264648156/HyDaP.
PDQ [13]	Distance-based; via PAM; Minimization of inverse cluster size, object-cluster likelihood, and weighted hybrid distance (Euclidean, normalized Manhattan, mismatch) based on variable type proportions; Centers' difference between successive iterations under threshold; `PDQ` in `FPDclustering` package [14].

represent the conditional probability of observing the category $\tilde{w}$ given the true category w, with columns of $\boldsymbol{\Theta}$ summing up to one. The unobserved and observed probabilities are linked by

$$\gamma_{\tilde{w}_l} = P(\tilde{W} = \tilde{w}_l) = \sum_{l=1}^{L} P(\tilde{W} = \tilde{w}_l | W = w_l) P(W = w_l) = \sum_{l=1}^{L} \theta_{\tilde{w}_l | w_l} \pi_l; \quad (1)$$

therefore, $\boldsymbol{\gamma} = \boldsymbol{\Theta}\boldsymbol{\pi}$ and $\boldsymbol{\pi} = \boldsymbol{\Lambda}\boldsymbol{\gamma}$, where $\boldsymbol{\Theta}$ is a matrix of misclassification probabilities with $\theta_{\tilde{w}_l|w_1}, \ldots, \theta_{\tilde{w}_l|w_L}$ in the $\tilde{w}_l$th row. Similarly, $\boldsymbol{\Lambda}$ is a matrix of reclassification probabilities, using $\lambda_{w|\tilde{w}} = P(W = w|\tilde{W} = \tilde{w})$.

4 Simulation Study

First, data without ME and Mi were generated separately for each cluster by means of the `genOrdNor` function from the `OrdNor` R package [16], which simulates a dataset with ordinal and normal variables having a predefined correlation matrix and marginals. Ordinal variables are generated by discretization of variables sampled from a multivariate normal distribution, and the ordinal components are treated as categorical.

For the generated datasets, $K = 2$, and $H = M = 2$ (therefore $P = 4$), $L_{1,2} = 2$. Marginal probabilities for the main category in each categorical variable, i.e., the most prevalent level within the variable, is 85%; each cluster has its own main category. Thus both categorical variables are informative of the clusters' structure. Analogously, interval scale variables carry information about the underlying clustering structure, since the cluster means were set at $\mu_1 = (-5, 0)$ and $\mu_2 = (0, 5)$, respectively. The total sample size N varies in $\{100, 500, 1000\}$ and clusters can have equal or different size. In case of different sizes, one cluster is 30% smaller than the other. The considered correlation between a numerical and a categorical variable is a special case of the Pearson's correlation coefficient [15], ranging from -1 to 1 as well. In this study, the correlation ρ varies in $\{0.15, 0.45, 0.60\}$. For each parameter combination, we conducted $R = 1000$ replications. Where applicable, 10 initializations have been used.

The original data were subsequently perturbed, adopting the ME and Mi models described in Sect. 3. For continuous variables, $U \sim \mathcal{N}_H(\mathbf{0}_H, \sigma^2 \mathbf{I}_H)$, where $\mathbf{I}_H$ denotes identity matrix and $\sigma \in \{0.5, 1, 2\}$ to simulate varying magnitudes of ME (low, medium, high). Analogously, the diagonal elements of the misclassification matrices varied in $\{0.1, 0.3, 0.5\}$ for obtaining three Mi levels. For each generated original dataset, there were nine *perturbed* datasets, given by the combinations of the ME and Mi levels (low-low, medium-low, high-low, etc.).

The Adjusted Rand Index (ARI) [17] was used to evaluate the clustering results. The ARI was computed for both the original and perturbed datasets, then the difference $\Delta ARI := ARI^{OR} - ARI^{MEM}$ was used to quantify the net impact of the perturbations on the clustering algorithm's ability to correctly identify the underlying structure. Assuming $ARI \in [0, 1]$, if the average $\Delta ARI > 0$ over the R replicates, then perturbation generally introduces a deterioration in clustering quality relative to the original scenario. If the average $\Delta ARI \leq 0$ the clustering quality either improves (< 0) or remains unaffected ($= 0$) despite the perturbation.

5 Results

Only results for equal-size clusters and $N = 500$ are shown, as outcomes are consistent across various cluster sizes and N values. The effects of correlation and perturbations on clustering results are reported in Fig. 1.

For k-prototypes at a low correlation level, increases in ME and Mi typically worsen clustering quality upon perturbation. The M-S algorithm demonstrates

similar trends. In contrast, HyDaP's clustering results are more significantly affected by increases in ME than Mi, yet the algorithm consistently achieves more accurate partitions close to true labels in the original scenario compared to perturbed scenarios. At medium to high correlation levels, k-prototypes exhibit the same qualitative behavior with increasing Mi levels, degrading clustering quality as in the low-correlation scenario. However, increasing ME levels tend to reduce the average ΔARI. Interestingly, at medium error levels and mostly for medium correlation, perturbed and original cases often yield similar results, a pattern also observed with the M-S algorithm. HyDaP displays consistent behavior across different correlation levels. The KAMILA and PDQ algorithms show less susceptibility to both perturbations and correlation levels. Higher correlations slightly increase the differences in results, regardless of the perturbation level, marking an exception in their generally stable performance.

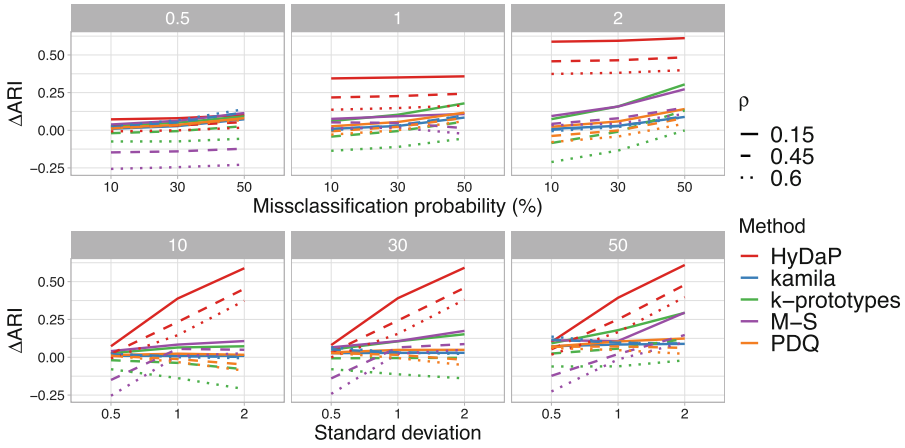


Fig. 1. Average ΔARI varying the ME standard deviation (plots in the top figure) and varying the Mi probabilities (plots in the bottom figure). Displayed results are for $N = 500$ and equal clusters' sizes. Parameters: $K = 2$, $H = M = 2$, $L_{1,2} = 2$.

6 Conclusion

This study reveals that the impact of correlation and perturbations on clustering outcomes is significantly influenced by the choice of algorithm, demonstrating that responses to these factors are highly method-dependent. Notably, there is no overarching trend across the evaluated algorithms, underscoring the complexity of clustering mixed-type data in the presence of errors and correlations. Algorithms frequently misidentify true cluster structures under diverse conditions, complicating the determination of generally unreliable scenarios. However, with low/medium error and low correlation levels, all algorithms except HyDaP demonstrated reliable results, implying their suitability for decision-making based on cluster analysis outcomes. Even algorithms with seemingly

similar conceptual underpinnings, such as PDQ and HyDaP, exhibit distinct behaviors under equivalent conditions.

Despite the results observed depend on the specific simulation settings used, our findings suggest that practitioners must carefully consider the specific characteristics of their data, including the extent of correlation and potential errors, when selecting a clustering algorithm. The varying resilience of k-prototypes, M-S, KAMILA, HyDaP, and PDQ to ME and Mi indicates that there is no one-size-fits-all solution for clustering mixed-type data in presence of correlation.

References

1. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley, Hoboken (2011)
2. Foss, A.H., Markatou, M., Ray, B.: Distance metrics and clustering methods for mixed-type data. Int. Stat. Rev. **87**(1), 80–109 (2019)
3. Buonaccorsi, J.P.: Measurement Error: Models, Methods, and Applications. CRC Press, Boca Raton (2010)
4. Tortora, C., McNicholas, P.D., Palumbo, F.: A probabilistic distance clustering algorithm using gaussian and student-t multivariate density distributions. SN Comput. Sci. **1**, 1–22 (2020)
5. Storlie, C.B., et al.: Clustering and variable selection in the presence of mixed variable types and missing data. Stat. Med. **37**(19), 2884–2899 (2018)
6. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Disc. **2**(3), 283–304 (1998)
7. Szepannek, G.: ClustMixType: user-friendly clustering of mixed-type data in R. R J. **10**(2), 200 (2018)
8. Modha, D.S., Spangler, W.S.: Feature weighting in k-means clustering. Mach. Learn. **52**, 217–237 (2003)
9. Foss, A.H., Markatou, M.: kamila: clustering mixed-type data in R and Hadoop. J. Stat. Softw. **83**, 1–44 (2018)
10. Foss, A., Markatou, M., Ray, B., Heching, A.: A semiparametric method for clustering mixed data. Mach. Learn. **105**(3), 419–458 (2016). https://doi.org/10.1007/s10994-016-5575-7
11. Wang, S., Yabes, J.G., Chang, C.C.H.: Hybrid density-and partition-based clustering algorithm for data with mixed-type variables. J. Data Sci. **19**(1), 15–36 (2021)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, Hoboken (2009)
13. Tortora, C., Palumbo, F.: Clustering mixed-type data using a probabilistic distance algorithm. Appl. Soft Comput. **130**, 109704 (2022)
14. Tortora, C., Vidales, N., Palumbo, F., Kalra, T., McNicholas, P.D.: FPDclustering: PD-Clustering and Factor PD-Clustering. R package version 2.2 (2022). https://CRAN.R-project.org/package=FPDclustering
15. Demirtas, H., Yavuz, Y.: Concurrent generation of ordinal and normal data. J. Biopharm. Stat. **25**(4), 635–650 (2015)
16. Amatya, A., Demirtas, H.: OrdNor: an R package for concurrent generation of correlated ordinal and normal data. J. Stat. Softw. **68**, 1–14 (2015)
17. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

From Bachelor to Master Degree: A First Sight of STEM Graduates' Choices

Giovanni Boscaino(✉) and Vincenzo Giuseppe Genova

Department of Economics, Business, and Statistics, University of Palermo, Palermo, Italy
{giovanni.boscaino,vincenzogiuseppe.genova}@unipa.it

Abstract. In Italy, as in much of the world, the presence of women in STEM disciplines and jobs is chronically lower than that of men. Many countries are taking action to encourage women to pursue studies and careers in those fields, adopting ad hoc policies. The study focuses on the decision-pathway of a STEM bachelor's degree in Italy. The data concern individual data of two cohorts of matriculates (2016 and 2018), relating to pre- and post-pandemic graduates, in order to take into account, as a possible determining factor, the social context of reference. The primary questions addressed are whether graduates opt to pursue further studies and, if so, whether they choose to follow the STEM pathway and remain at the same institution. In addition, we also focus on the possible association between the choices made and the gender of the student. Preliminary results suggested by a Multiple Correspondence Analysis approach seem to indicate that there is no effect associated with the pandemic crisis but that the choices seem to be partly dictated by gender, academic success, geographical area, and the type of diploma held (more or less consistent with a STEM educational path).

Keywords: STEM · Italy · Women · MCA · higher education

1 Introduction

In Italy, the under-representation of women in STEM fields remains a stubborn problem, even though interest in these areas has been on the rise over the past ten years. The Global Gender Gap Report for 2023 [1] underscored that Italy's overall ranking slipped from 63rd in 2022 to 79th out of 146 nations in 2023, largely due to women's participation in politics. Nonetheless, there was a minor uptick in terms of economic participation and opportunities.

The 2018 ISTAT report revealed that only 20% of STEM graduates in Italy are women, a significantly lower figure than the European average of approximately 40% [2]. Numerous investigations have attempted to uncover the reasons behind this disparity. Some point to sociocultural elements such as gender stereotypes and societal expectations [3], while others underscore the role of education and school guidance [4]. However, most of these studies rely on aggregate data at

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 111–116, 2025.
https://doi.org/10.1007/978-3-031-64346-0_19

the national or regional level, which can obscure significant local variations [5]. Furthermore, most recent Italian literature focused mostly on High School students' performance [6], on University students about staying in their program, dropping out, or switching to change fields of study [7] or due to mobility choices [8]. Using administrative microdata from the Italian Ministry of Universities and Research and focusing on students enrolled in a STEM field from 2010 to 2015, our aim is twofold: to identify the specific patterns that differentiate students in terms of staying in their program, dropping out, or changing fields and focus on gender differences. Only a handful of studies have employed sophisticated statistical techniques to examine the gender gap in Italy in STEM [2]. In our study we refer to the data processed in accordance with the Research Protocol for the Study "From high school to the job placement: analysis of university careers and University mobility from Southern to Northern Italy" among the Ministry of University and Research, the Ministry of Education and Merit, the University of Palermo as the lead institution, and the INVALSI Institute. The reference researcher is Massimo Attanasio. The data encompasses comprehensive information about every student enrolled at a public university in Italy since 2008. For students enrolled at the University of Palermo, the data are linked to the results of the AlmaLaurea annual survey on university experience and job placement of graduates. With access to this extensive database, in this paper we concentrate on the choices related to the completion of a STEM bachelor's degree. The primary questions addressed are whether graduates opt to pursue further studies and, if so, whether they choose to follow a STEM pathway and remain at the same institution. In addition, we also focus on the possible association between the choices made and the gender of the student. The objective is to offer an initial response that can be used to guide future research. In this paper, we investigate the phenomenon under study from a descriptive and a graphical point of view, thanks to the use of Multiple Correspondence Analysis, which should allow us to discern characteristics of the students considered.

2 Data

The study considered the 2016 and 2018 cohorts of pure enrollees in Italian non-telematic universities in three-year degree programmes in the STEM areas, who graduated within four years. The choice also concerned these cohorts because we want to investigate the possible effect of the restrictions due to the COVID emergency that started in January 2020 and 'ended' in 2022.

Figure 1 depicts the details of STEM enrolments in Italy during the academic year 2016/17. Upon closer analysis, approximately 61% of STEM graduates in Italy are male students, and this percentage remains consistent even in the academic year 2018/19. Regarding the progression to master's studies, 89% of women choose to pursue a master's degree and, of these, 24% opt for a different university than the one where they got their bachelor's degree. For male students, the percentage continuing to a master's degree remains the same, but 16% decide to change universities. However, in the academic year 2018/19, there

is a significant increase in the percentage of students who do not continue their studies, rising from 11% in the academic year 2016/17 to 36.7%. No significant gender-based differences are observed in terms of non-prosecution of the studies. Furthermore, regardless of the year considered and gender, nearly all STEM graduates choose to pursue a master's degree in STEM fields.

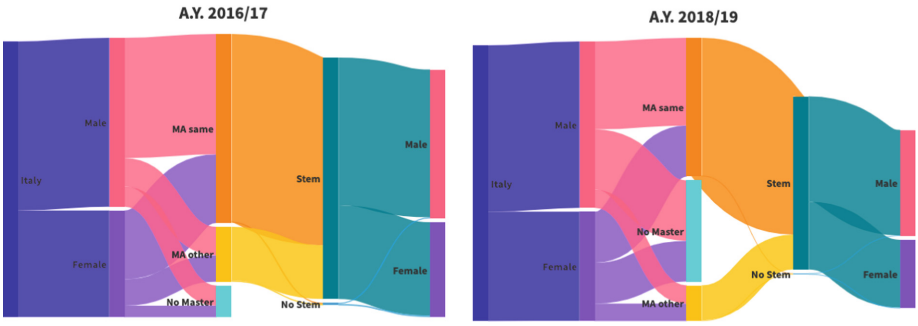


Fig. 1. Analysis of STEM student enrolment in Italy

3 Methodology

In this preliminary work, we refer to multiple correspondence analysis (MCA) as an exploratory approach to investigate the association structure among among those variables linked to the graduates' choices. MCA, in brief, is a statistical technique similar to Factorial Analysis but on qualitative variables. Beyond the factorial structure, MCA provides also a graphical representation of relationships among multiple categorical variables. As an extension of simple Correspondence Analysis, MCA transforms categorical variables into a set of continuous variables, which are then represented in a multidimensional space [9], where each dimension represents a common latent factor. The resulting factors emerge in an order that reflects the amount of data variability each one can capture. In this way, each subsequent factor explains a progressively smaller portion of the total variability of the data [10].

With respect to the usual definitions related to this method, we recall just some basic concepts [11]:

- **Active Variables:** They are the main variables used in the MCA. They influence the results of the analysis, including the computation of dimensions, eigenvalues, and factor loadings.
- **Illustrative Variables**: They are not used in the computation of the dimensions, but their coordinates are predicted based on the results obtained from the active variables. This allows you to interpret these variables within the context of the structure revealed by the active variables, without them influencing that structure.

- **Inertia**: In MCA, inertia is a measure of the total variance in the data set. It is calculated as the weighted sum of the chi-square distances between the categories. The inertia of a category on a given axis is the amount of variance that this category contributes to this axis. The total inertia will be the sum of the inertias of all categories. The higher the inertia, the more the data is dispersed.
- **Greenacre's Correction**: Inertia can be influenced by the number of categories in the variables. Greenacre's correction modifies the inertia so that it is not influenced by the number of categories. This makes inertia a more reliable indicator of the variance explained by each dimension.

The interpretation of the MCA results involves examining the proximity of the categories in the multidimensional space. Categories that are close to each other in space are similar to each other.

MCA has found extensive applications in various fields such as sociology, marketing, and medicine, where it helps to understand complex multivariate relationships among variables. Despite its wide usage, the interpretation of MCA results should be approached with caution, especially when dealing with large datasets with numerous variables.

4 Preliminary Results and Comments

After having depicted the decision-pathway of STEM Bachelors' degree in a descriptive manner, we investigated the graduates choices via the Multiple Correspondence Analysis. As active variables, we have considered:

- Graduate's decision: enrolling in a STEM master's degree course (STEM), enrolling in a non-STEM master's degree course (NO STEM), non-enrolment in university (NO);
- School diploma mark, in classes: < 70, [70, 90], > 90 (in the Fig. 2: $D < 70$, $D70 - 90$, $D > 90$, respectively). The classes are built to identify convenient different school performances;
- Graduation mark, in classes: [60, 89], [90, 99], [100, 104], and > 104 (in the Fig. 2: $L < 90$, L90-99, L100-104, $L >= 105$, respectively). The classes are built to identify convenient different graduates performances;
- Geographical macro-area of the university of enrolment for the Master's degree course: North East, North West, Centre, South, Isles. We have followed the ISTAT 2024 classification of the Italian Regions.

As illustrative variable, we have considered just the Gender (Male, Female). In addition, we have also considered the geographical macro-area of the university of enrolment for the Bachelor's degree course, but this was highly correlated with that of the Master's degree (Cramer's V = 0.92) and therefore we did not consider it in the analysis. Figure 2 is for the cohort of enrollees in 2016 and represents the factorial plan of the first two dimensions, which together catch a Greenacre's inertia of nearly 55%. The figure suggests that the choice not

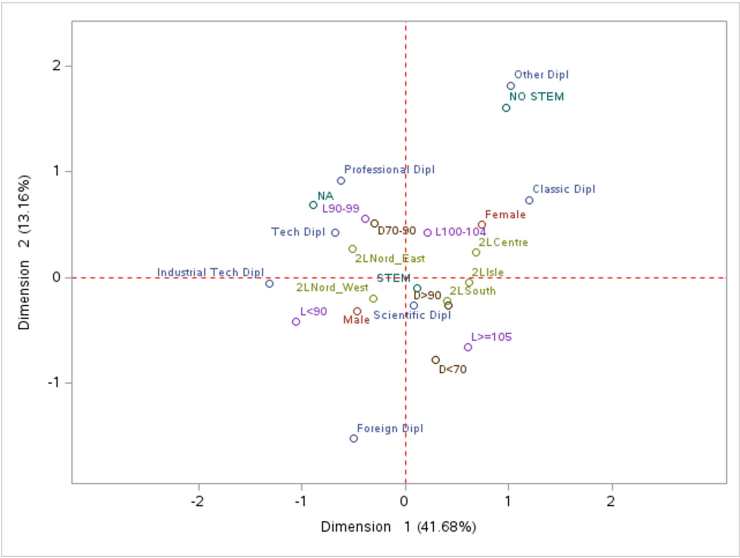


Fig. 2. MCA Projections on the first 2 dimensions, Students' cohort 2016 (Greenacre's adjusted inertia percentages are reported in parentheses)

to continue with a STEM degree is more associated with profiles linked to the female gender, a good degree grade, regions of the south-central, possession of a classical maturity diploma, or "other". On the other hand, a clear picture does not emerge, probably because the choice to continue on a STEM path is the most undertaken. As reported Sect. 2, data highlight most of STEM graduates decides to enroll in a STEM Master Degree Course. In terms of a cloud of points, we notice a proximity with the male gender, with the geographical area of the North, possessing a scientific maturity diploma title and a decent degree grade. Finally, the non-prosecution of studies in a Master Degree seems more associated with those who achieve a degree grade that is at most sufficient and those who possess a technical or professional maturity diploma and, in this case, from the North. The arrangement of the active variables on the plane and the analysis suggests that the two dimensions can be mainly associated with the degree grade, the geographical area of registration (Dimension 1), and the type of diploma obtained (Dimension 2). For brevity, the graph related to the cohort of enrollees in 2018 is not reported. Also, the explained inertia is greater than 55% in this case. The identified dimensions remain the same as the previous ones. As for the possible profiles, in this case, the difference is found for those who decide to continue on a STEM path. This choice is indeed associated with a high degree grade, a geographical region of the south-central, and always with the male gender.

Aknowledgement. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union - NextGenerationEU - Project Title Stem in Higher Education & Women INequalitieS [SHE WINS], CUP I53D23004810006, Grant Assignment Decree No. 1060 adopted on 07/17/2023 by the Italian Ministry of University and Research (MUR).

This study was partially funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 - CUP B73C22001260006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

1. World Economic Forum: The Global Gender Gap Report, Insight Report June 2023. Geneva: World Economic Forum (2023)
2. ISTAT: Donne e uomini in Italia. Roma: Istituto Nazionale di Statistica (2018)
3. Nosek, B.A., Smyth, F.L., Sriram, N., et al.: National differences in gender-science stereotypes predict national sex differences in science and math achievement. In: Proceedings of the National Academy of Sciences, vol. 106. no. 26, pp. 10593–10597 (2009)
4. OECD: The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence. Paris: OECD Publishing (2015)
5. Blickenstaff, J.C.: Women and science careers: leaky pipeline or gender filter? Gend. Educ. **17**(4), 369–386 (2005)
6. Priulla, A., D'Angelo, N., Attanasio, M.: Gender differences in stem courses: analysis of Italian students' performance. In: Book of Abstracts of the International Conference of the journal Scuola Democratica. Reinventing Education, Rome, Associazione "Per Scuola Democratica" (2021)
7. Tocchioni, V., Galluccio, C., Morabito, M.F., Petrucci, A.: Students enrolled in STEM discipline in Italy: patterns of retention, dropout and switch. In Book of the short paper SIS 2022. Pearson (2022)
8. D'Agostino, A., Ghellini, G., Longobardi, S.: Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. Electron. J. Appl. Stat. Anal. **12**(4), 826–845 (2019)
9. Kassambara, A.: Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra. Vol. 2. Sthda (2017)
10. Di Franco, G.: Multiple correspondence analysis: one only or several techniques? Qual. Quant. **50**, 1299–1315 (2016)
11. Greenacre, M.J., Blasius, J.: Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, Boca Raton (2006)

Statistical Models for Health Monitoring of Rare Events in Railway Tracks

Nikolaus Haselgruber[1(✉)] and Ingolf Nerlich[2]

[1] CIS consulting in industrial statistics GmbH, 4170 Haslach, Austria
nh@cis-on.com
[2] Schweizerische Bundesbahnen AG, 3000 Bern, Switzerland
ingolf.nerlich@sbb.ch

Abstract. Squats are punctual material failures at railway tracks which can lead to critical effects when not detected or removed in time. This paper proposes a statistical modelling approach to describe the relation between risk of squat occurrence and potential root causes, such as track characteristics, traffic patterns and load data, considering multicollinearity and rare events. Data from the Swiss railway were used to develop models for track health monitoring.

Keywords: Squats · rare events · health monitoring · railway tracks

1 Situation

Railway reliability is influenced by factors such as trains, human elements, traffic control but also by track infrastructure. Squats, or track material failures, can lead to serious risks if not promptly addressed. Despite ongoing research (c.f., [1–4]), the genesis of squats and their distribution across the rail network is not fully understood yet. Variations in track material, train technology, operational and maintenance pattern are suspected to influence squat formation.

Given their infrequent occurrence compared to the network's total length, this paper aims to explain the squat risk using Swiss railway (SBB) data. After data aspects, a statistical model for squat risks and its application is discussed, followed by implications for future research.

2 Data Generation and Preparation

The Swiss railway net is digitalized and resolved in equi-distant positional points every meter, separated for both left and right track [5]. An overall length of approx. 5200 km results in 10.4 million different positional points. Besides spatial position, tracking squats involves also recording temporal status and leads to either track replacement or grinding for removal. An *observation* represents a spatio-temporal track segment, identified by geographical coordinates and time interval, either bounded by track replacement/treatment or squat detection dates. Squat presence is marked as a binary outcome for each segment.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 117–122, 2025.
https://doi.org/10.1007/978-3-031-64346-0_20

Analysis integrates various data types, including track characteristics (e.g., gauge, material, track shape), load data (e.g., traffic volume, train types), and events (e.g., maintenance actions, squat detection), resulting in > 700 variables.

3 Squat Modelling

3.1 Modelling Aspects

To develop reasonable stochastic models for the description of the relation between squats and its potential root causes, several aspects have to be considered:

(a) Squat occurrence as binary outcome can be modelled, e.g., by logistic regression or as accelerated lifetimes with the response as censoring indicator.
(b) The proportion of squats is small, thus, *case-control* sampling (c.f., e.g., [6]) of failure-free observations may be promising.
(c) The large set of potential regressors is partially correlated, imbalanced and contains many missing values. This can lead to instable results. Pre-selection by multivariate analyses and checking multi-collinearity will be required.
(d) Implementation of a track health monitoring primarily requires the understanding of the root causes of squats to highlight potentially critical segments. A second priority would be the estimation of time to failure.
(e) International comparison showed differences in data quality and depth between rail operators [7]. A model for broad applicability should prefer easily available and comparable regressors, such as basic track characteristics.

To address these items, we propose a generalized linear model, choosing regressors with input from field experts without strict criteria for balance or distribution but checking multi-collinearity. With focus on the low incidence of squats, we incorporate case-control sample bootstrapping for non-failure observations.

3.2 Model Formulation

Let $\pi_i = P(Y_i = 1)$ denote the probability of squat occurrence and $x_{i1}, \ldots, x_{iK}$ values of potential regressors for $i = 1, \ldots, n$. To model effects of higher order or interactions, x_{ik} can alternatively represent a continuous functional transformation or be composed of several regressors. $y_i \in \{0, 1\}$ are realizations of stochastically independent random variables Y_i, following a *Bernoulli* distribution with success probability $\pi_i \in [0, 1]$. To avoid model-based predictions outside this range, the nonlinear transformation $Logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$ serves as response variable which is described by the *linear predictor* $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK}$ with the unknown regression coefficients $\beta_0, \ldots, \beta_K$. Consequently,

$$P(Y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)} = F(\eta_i),$$

which is the cumulative distribution function of the *logistic* distribution and $Logit(\pi_i) = F^{-1}(\pi_i) = \eta_i$. Thus, $\exp(\eta_i) = \frac{\pi_i}{1-\pi_i}$ which is called the *odds,* i.e. the ratio of failure to failure-free probability. Furthermore, the impact of a change in a regressor variable $\Delta x_{ik} = x_{ik,\mathbf{new}} - x_{ik,\mathbf{initial}}$ can be expressed as

$$\frac{\exp(\beta_0 + \ldots \beta_{k-1}x_{i,k-1} + \beta_k x_{ik,\mathbf{new}} + \beta_{k+1}x_{i,k+1} + \ldots)}{\exp(\beta_0 + \ldots + \beta_{k-1}x_{i,k-1} + \beta_k x_{ik,\mathbf{initial}} + \beta_{k+1}x_{i,k+1} + \ldots)} = \exp(\beta_k \Delta x_{ik})$$

and is called the *odds ratio,* i.e., the change in the ratio of failure to failure-free probability depending on Δx_{ik}. Multi-collinearity can bias the interpretion of single effects. It is controlled by checking the variance inflation factors [8].

3.3 Parameter Estimation

Maximization of the likelihood as a function of unkown parameters $\beta_0, \ldots, \beta_K$

$$L(\beta_0, \ldots, \beta_K) = \prod_{i=1}^{n} \left(\frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right)^{y_i} \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(\eta_i)} \right)^{1-y_i},$$

provides the estimates $\hat{\beta}_0, \ldots, \hat{\beta}_K$ for observed and fixed $(y_i, x_{i1}, \ldots, x_{iK})$. The estimated failure probability is $\hat{\pi}_i = F(\hat{\eta}_i = \sum_k \hat{\beta}_k x_{ik})$ with $x_{i,1} = 1$. To balance between model accuracy and complexity, stepwise model selection is applied. We use the Akaike information criterion $AIC = -2 \ln L(\hat{\beta}_0, \ldots, \hat{\beta}_K) + 2(K+1)$ and select backwards, starting from the full model, that set of parameters which provides the best compromise, i.e., the minimal AIC.

3.4 Model Diagnosis and Validation

Diagnosis is done by residual and leverage plots. A 2-fold cross-validation, where $\hat{\pi}_i = \exp(\hat{\eta}_i)/(1 + \exp(\hat{\eta}_i))$ is compared with a threshold $\gamma \in [0; 1]$, provides the decomposition $n = n_{TP}(\gamma) + n_{TN}(\gamma) + n_{FP}(\gamma) + n_{FN}(\gamma)$ in *true positive,* i.e., correctly classified squats, *true negative,* i.e., correctly classified non-squats, as well as *false positive* (erroneously classified squats) and *false negative* (erroneously classified non-squats) cases. The model is assessed by $AUROC \in [0, 1]$, the area under the receiver operating characteristic curve, derived from true positive portion or *sensitivity* $p_{TP}(\gamma) = n_{TP}(\gamma)/(n_{TP}(\gamma) + n_{FN}(\gamma))$, over the false positive portion or *1-specificity* $p_{FP}(\gamma) = n_{FP}(\gamma)/(n_{FP}(\gamma) + n_{TN}(\gamma))$. Higher values are better, values ≤ 0.5 indicate worthless models (c.f., [9]).

3.5 Case-Control Sample Bootstrapping

To overcome potential problems caused by rare occurrence of squats, we propose case-control sample bootstrapping as follows:

1. Determine the number of squats in the full sample as $n_f = \sum_i y_i$
2. Repeat the following two steps 1000 times, $j = 1, \ldots, 1000$:

3. Draw a random sample of size $n_0 = 2n_f$ from the non-squat observations
4. Carry out stepwise estimation. Report AIC_j, $AUROC_j$, $\hat{\beta}_{0j}, \ldots, \hat{\beta}_{Kj}$, set effects of excluded regressors to 0 and their p-values to 1.
5. Summarize distributions of effects and p-values, build averages $\bar{\hat{\beta}}_0, \ldots, \bar{\hat{\beta}}_K$
6. Select that model j^* which provides $j^* = j : \sum_k (\hat{\beta}_{kj} - \bar{\hat{\beta}}_{kj})^2 \rightarrow \min!$

Although sampling with $n_0 = n_f$ would provide better balance in the response, we prefer a higher sample size at a still reasonable balance. To correct the bias in $\hat{\pi}$ caused by case-control sampling, we consider the estimate $\exp(\hat{\beta}_{0F})$ of the odds under the null model $\eta_i = \beta_0$ with the full sample, where $\hat{\pi} = n_f/n$ and the estimate $\exp(\hat{\beta}_{0C})$ correspondingly for the case-control sample with $n_0 = cn_f$ and $1 \leq c \leq (n - n_f)/n_f$, then it is $\exp(\hat{\beta}_{0F} - \hat{\beta}_{0C}) = c\hat{\pi}/(1 - \hat{\pi})$, which allows to estimate $\hat{\beta}_{0F}$ from the case-control sample as $\hat{\beta_{0C}} + \ln(c\hat{\pi}/(1-\hat{\pi}))$.

3.6 A Synthetic Case to Study the Approach

Assume the relation $\eta = -8 - x_1 + x_2 - x_2 x_2$ to study the proposed approach before applying it to real data. For each combination of the full 2^2-factorial design of (x_1, x_2) over $\{-1, 1\} \times \{-1, 1\}$, we simulate 100,000 Bernoulli realizations y with $\pi_i = \exp(\eta_i)/(1 + \exp(\eta_i))$, providing $n = 400,000$ and $n_f = 647$, i.e., $\hat{\pi} \approx 0.16\%$. The `R` [10] function `glm(y~x1*x2,family=binomial(link="logit"))` applied to the full sample and to case-controlled samples of size $3n_f = 1941$, both times without stepwise model selection, provides the results as shown in table 1.

Table 1. GLM fit results of simulated rare event data with orthogonal design

Par.	Full sample ($n = 400,000$)				Selected case-controlled sample ($n = 1941$)			
	Est.	Std. err.	z value	$P(>\|z\|)$	Est.	Std. err.	z value	$P(>\|z\|)$
Int.	−8.2854	0.1506	−55.026	< 2e−16	−2.5533	0.1531	−16.6769	< 2e−16
x_1	−0.8864	0.1506	−5.887	3.94e−09	−0.8848	0.1531	−5.7793	7.75e−09
x_2	1.2330	0.1506	8.188	2.65e−16	1.2330	0.1531	8.0531	8.07e−16
x_1:x_2	−1.0891	0.1506	−7.233	4.72e−13	−1.0845	0.1531	−7.0833	1.41e−12

Sampling has negligible impact on the results, except for β_0, where bias correction leads to $-8.2854 = -2.5533 + \ln(2(647/400,000)/(1 - 647/400,000))$. Coefficients of variation of the estimates among the case-controlled samples are $\approx 2.5\%$, all p-values are <1e-08, i.e., results are robust for the orthogonal design. Now, a further regressor is added as $x_3 = -1$ for all except two cases, one with $y = 0$ and one with $y = 1$ where $x_3 = 1$, but identical settings of x_1 and x_2. Table 2 shows the corresponding results. While, based on the full sample, x_3 is recognized as significant effect, the case-controlled sample bootstrap unmasks the multi-collinearity with the intercept very clearly. Since the AIC-conducted stepwise regression does not exclude x_3 automatically, manual review and model adaptation is required and leads iteratively back to the results shown in table 1.

Table 2. GLM fit results of simulated rare event with non-orthogonal design

Par.	Entire sample ($n = 400,000$)				Selected case-controlled sample ($n = 1941$)			
	Est.	Std. err.	z value	$P(>\|z\|)$	Est.	Std. err.	z value	$P(>\|z\|)$
Int.	−3.3793	0.7244	−4.665	3.09e−06	5.7417	162.3719	0.0353	9.71e−01
x_1	−0.8408	0.1573	−5.344	3.94e−09	−0.8566	0.1598	−5.3615	8.25e−08
x_2	1.2785	0.1573	8.126	9.10e−08	1.2566	0.1598	7.8649	3.69e−15
x_3	4.9517	0.7416	6.677	2.44e−11	8.3401	162.3720	0.0514	9.59e−01
$x_1{:}x_2$	−1.1347	0.1573	−7.212	5.52e−13	−1.1158	0.15978	−6.9836	2.88e−12

4 Model Application to SBB Squat Data

Initially, we consider all squat types in a full sample with $n = 1,502,670$ and $n_f = 11,800$, i.e., $\hat{\pi} = 0.79\%$, emphasizing easily obtainable regressors such as observation period duration, cargo transport proportion, accumulated transport weight per day and period, slope gradient, curvature, and track steel quality, along with some interactions. Pre-selection keeps variance inflation factors < 2.5 for single and < 10 for combined effects, i.e., correlations among regressors are small, except values ≈ 0.5 for weights and days, so multi-collinearity is diminished. Table 3 shows significant impact for most regressors: E.g, low cargo transport proportion, low curvature, R350 steel quality, a longer period combined with lower accumulated weight and a shallow slope gradient, contribute to squat risk. Bias correction of case-controlled sampling with $c = 2$ provides $\beta_{0F} = \beta_{0C} + \ln(2\hat{\pi}/(1-\hat{\pi})) \approx -4.549$. E.g., a change in the proportion of cargo transport from 100% to 0% leads to $\Delta x_1 = -1$ and to an odds ratio of $\exp(2.57) \approx 13$. While achieving $AUROC \approx 0.71$, more complex models tailored to specific squat types and incorporating advanced load calculations (c.f., [11]) can attain $AUROC > 0.9$, offering substantially improved squat modeling capabilities.

Table 3. GLM fit results of SBB squat data: all squats/basic parameters

Parameter		Unit	Est.	Std. Err.	z value	$P(>\|z\|)$	
Intercept			−4.032e−01	4.407e−02	−9.900	< 2e−16	***
Proportion cargo	x_1	[1]	−2.570e 00	1.019e−01	−25.226	< 2e−16	***
Acc. weight/day	x_2	$[t]$	7.916e−07	1.030e−06	0.796	0.4421	
Acc. weight/obs. period	x_3	$[Mt]$	−1.195e−02	2.098e−03	−5.697	1.22e−08	***
Days in observation period	x_4	$[d]$	4.720e−05	2.787e−05	1.693	0.0904	.
Absolute slope gradient	x_5	[‰]	−9.923e−03	2.123e−03	−4.666	3.07e−06	***
Curve radius > 5 km [TRUE]	x_6	[-]	0.636e−01	2.674e−02	23.782	< 2e−16	***
Steel quality [R350]	x_7	[-]	7.275e−01	2.942e−02	24.730	< 2e−16	***
Simple thermal load index	x_8	[-]	1.438e−01	5.749e−02	2.501	0.0124	*
Interaction 1	$x_1{:}x_2$	$[t]$	−1.719e−05	3.248e−06	−5.291	1.22e−07	***
Interaction 2	$x_3{:}x_4$	$[dMt]$	2.577e−06	1.066e−06	2.418	0.0156	*

5 Summary and Outlook

In this paper we discuss the challenges of rare event modeling for railway track failures, proposing and applying case-control sample bootstrapping to data from the Swiss railway. Significant effects such as curvature, slope gradient, and track steel quality have been identified. However, extensive effort is needed for data collection and preparation, particularly in detecting, describing, and classifying squats. Based on current insights, more sophisticated measurements and load calculations may enhance model quality and predictive accuracy. Future research includes extension to Austrian (OEBB) and German (DB) railway data and exploring lifetime models to further improve track health monitoring capabilities.

Acknowledgement. The publication was written in cooperation with Virtual Vehicle Research GmbH in Graz and partially funded within the COMET K2 Competence Centers for Excellent Technologies from the Austrian Federal Ministries for Climate Action (BMK) and Labour and Economy (BMAW), the Province of Styria (Dept. 12) and the Styrian Business Promotion Agency (SFG). The Austrian Research Promotion Agency (FFG) has been authorised for the programme management.

References

1. Kerr, M., Wilson, A., Marich, S.: The epidemiology of squats and related rail defects. In: Conference on railway engineering (2008)
2. Luther M., Heyder R., Mädler K.: Prevention of multiple squats and rail maintenance measures. In: 11th International Conference on Contact Mechanics and Wear of Rail/wheel Systems (CM2018) Delft, the Netherlands (2018)
3. Muhamedsalih, Y., Hawksbee, S., Tucker, G., Stow, J., Burstow, M.: Squats on the Great Britain rail network: possible root causes and research recommendations. Int. J. Fatigue **149** (2021). https://doi.org/10.1016/j.ijfatigue.2021.106267
4. Schamberger, S.: Der Squat aus Sicht der OEBB, OEVG: Squats, University of Technology Graz (2021)
5. Nerlich I.: Squat - ein Gesamtsystemversagen! Den Ursachen mit netzweiter SBB-Statistik auf der Spur?, OEVG: Squats, University of Technology Graz (2021)
6. King, G., Zeng, L.: Logistic Regression in Rare Events Data. Published online by Cambridge University Press (2001). https://doi.org/10.1093/oxfordjournals.pan.a004868
7. Tucker, G., et al.: Statistical analysis of factors influencing squat formation and growth. In: Presentation at Squats22NL University of Delft (2022)
8. Fahrmeir, L., Kneib, T., Lang, S.: Regression. 2nd edition, Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01837-4
9. Agresti, A.: Categorial Data Analysis. 2nd edition, Wiley, New Jersey (2002). https://doi.org/10.1002/0471249688
10. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022). www.R-project.org/
11. Nerlich, I.: Netzweite statistische Analyse von Squat-Rollkontaktermüdungsfehlern unter Berücksichtigung von Kontaktgeometrie und Zusammensetzung der Traktionsmittel in einem Bahnsystem mit Mischverkehr. Submitted Dissertation, TU Berlin (2024)

Unconventional Data and Innovation: Are Innovative SMEs' Web Pages Different?

Carlo Bottai[3], Lisa Crosato[1], Josep Domenech[2], Marco Guerzoni[3], and Caterina Liberati[3(✉)]

[1] Università Ca' Foscari, Cannaregio 873, 30121 Venice, Italy
[2] Universitat Politècnica de València, Camí de Vera, s/n, 46022 Valencia, Spain
[3] Università di Milano-Bicocca, p.zza Ateneo Nuovo 1, 20126 Milan, Italy
caterina.liberati@unimib.it
https://www.unimib.it/caterina-liberati

Abstract. The work investigates the potential use of corporate websites as an alternative data source for research in innovation studies. Traditionally, research in this field relies on data gathered from financial statements, patents or surveys. Our study bridges the gap between standard economic data of firms and web-based data, contributing to the ongoing debate with an innovative approach based on unconventional data. Unlike the predominant focus in the existing literature on the linguistic content of a web page, we propose the usage of HTML tags too. The exploratory study has been conducted on a sample of Italian companies. The results support the hypothesis that innovative firms tend to build their corporate websites differently.

Keywords: Innovation · SMEs · HTML code · web-scraping

1 Conventional and Unconventional Data to Assess Innovation

Assessing the level and depth of innovative activity within a company is a complex task, mainly due to the intricate nature of innovation itself and the tendency for innovative practices to be embedded within a company's operations and the collective knowledge of its workforce. Over the past fifty years, operational research in the field of innovation has made significant progress in developing various methods and measures to understand this phenomenon. This assessment usually relies on patents surveys, and financial data as the main sources of information. However, each of these sources has its limitations in capturing hidden innovative elements, especially when dealing with Small and Medium-sized Enterprises (SMEs), which are vital for national economies.

Patents, despite being widely used in innovation research, have limitations such as differences in patenting rates among industries, the existence of technological knowledge that cannot be patented, and the tendency for patents to

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 123–127, 2025.
https://doi.org/10.1007/978-3-031-64346-0_21

protect inventive steps that may not always represent true innovations. Moreover, SMEs often lack the resources and infrastructure to engage in patenting and prefer other forms of intellectual property protection. As a result, patents may not accurately reflect the contributions of SMEs to the innovation process.

Financial statements, another common source of data in innovation studies, provide insights into innovative activities like research and development (R&D) spending and allow for inferences about a company's productivity, profitability, and growth. However, for SMEs, a significant portion of R&D costs may be informal and hidden within employee expenses rather than explicitly labeled as R&D spending in financial reports.

Surveys, while designed to collect diverse information, are prone to response biases and misunderstandings of questions by respondents. Furthermore, the EU Community Innovation Survey (CIS), which targets a wide range of European companies, occasionally includes questions about innovation processes but does not consistently survey small businesses, making it difficult to study these entities over time.

For all these reasons, it seems that conventional data sources are inadequate in fully capturing innovation's latent features [1,2]. Additionally, innovation policy indicators derived from these sources may not be sufficiently updated to reflect the current situation.

Hence, in this study we explore the idea to use corporate websites of SMEs, as manifestations of their activities, as an additional data source for constructing indicators of firms' innovative characteristics. Companies often design their websites as virtual platforms to promote their products and disseminate information pertinent to their operations, resulting in website content being closely linked to their economic activities [3]. Furthermore, corporate websites are publicly accessible and regularly updated, making them promising candidates for addressing some of the limitations associated with conventional data sources. Consequently, a portion of the academic literature has begun to utilize web scraping techniques for research purposes (e.g., [4,5]) and to leverage corporate websites for analyzing firms' innovative activities [7–13].

2 Data Description

Our study centers on Italian manufacturing SMEs operating in 2016. The constructed dataset integrates both traditional and innovative data sources. We compiled firm-level information from Orbis, Bureau van Dijk (BvD), including indicators such as the number of employees and total assets, as well as website urls (if available), along with comprehensive company details like location, identification number, business name, address, and finally two stratification variables as industrial sector (NACE) and geographical location (NUTS 2). Therefore, our data matrix was composed by 77,993 firms. In order to classify firms in terms of their innovativeness, we appended to the matrix further information about the presence of the company in the list of 'innovative' SMEs collected by the Italian Chamber of Commerce's Business Register in compliance with the Italian Startup Act (221/2012 law).

Unfortunately, the URLs provided by Orbis do not always consistently lead to the corporate websites of the specified businesses. Consequently, we implemented a screening process to mitigate the risk of potential discrepancies. Following the methodology outlined by Barcaroli et al. [14], our approach scrutinizes various pieces of information on the website's homepage (such as identification numbers, addresses, and postcodes) to validate them against the data obtained from Orbis[1]. This led to a reduction in the sample size, but it enhanced the quality of the data utilized in the study. Additionally, the web pages corresponding to validated websites were sourced from the Wayback Machine (https://archive.org/web/), an open internet archive of websites that enables users to track the evolution of many website over the years since its inception. We have accessed the 2016 archived version of each firm's website URL and retrieved the corporate homepage when present.

After applying these data checking/cleaning procedures, 43,335 SMEs remained. From this sample of firms, we gathered information on every HTML tag utilized to structure their respective web pages. Only HTML tags appearing more than three times throughout the entire document corpus were retained, resulting in a final set comprising 711 HTML tags and five aggregate web-based statistics (as in [4–6]).

3 Explorative Analysis of Innovative SMEs

Our sample included only 178 "innovative SMEs". To ensure robust comparisons, we paired the subset of innovative firms with 100 similar subsets based on size, region, and industry. To discern differences in the HTML structure between innovative and non-innovative corporate websites, we compared the aggregate statistics and the HTML tags collected on each of the non-innovative samples against the group of innovative SMEs.

Our study pursues two main objectives: initially, we examine whether there are differences in the aggregate statistics between innovative and non-innovative SMEs. Subsequently, we explore whether HTML tags gather into natural groupings leading to different websites' organizational structure corresponding to different types of firms. For brevity, we present only essential results from the first analysis.

We computed five aggregate statistics that assess various size aspects of a corporate website. The variables `html-size` and `text-size` measure the size (in bytes) of the HTML code and the length of text within each web page, respectively. The variable `images` indicates the number of images included in the document, while `hyperlinks` tallies the number of user-clickable hyperlinks present in the HTML document. Lastly, `stylesheets` consolidates the count of external resources, primarily CSS files, used in the web page.

Table 1 shows the median values of the p-values resulting from the paired t-tests and Wilcoxon signed-rank tests. Results reveal that the null hypothesis

[1] The detailed procedure is available in [15].

Table 1. Median p-value of Paired t-test and Wilcoxon signed-rank test

Size Variable	Paired t-test	Wilcoxon signed-rank
`html-size`	0.001	0.000
`text-size`	0.012	0.003
`images`	0.082	0.041
`hyperlinks`	0.008	0.002
`stylesheets`	0.000	0.000

of equivalence between innovative and non-innovative SMEs regarding the size measures is consistently rejected at the 5% significance level, but for the number of images (10% significance according to the paired t-test). Further research will refine these results and offer evidence regarding differences in terms of all the collected HTML tags, indicating distinct coding practices between innovative and non-innovative firms.

Concluding, corporate websites have shown potential for measuring SMEs' innovation. Until now, the text of these websites has been used. However, our results support the hypothesis that the HTML code of these websites too is informative of the innovativeness of a business. We hope this additional feature will be included in future web-based firm-level innovation indicators.

References

1. OECD.: Frascati Manual: The Proposed Standard Practice for Surveys of Research and Experimental Development. OECD Publishing (1963)
2. OECD.: Oslo Manual: OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. OECD Publishing (1992)
3. Domènech, J., de la Ossa, B., Pont, A., Gil, J.A., Martinez, M., Rubio, A..: An intelligent system for retrieving economic information from corporate websites. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 573–578 (2012)
4. Blázquez, D., Domènech, J., Debón, A.: Do corporate websites' changes reflect firms' survival? Online Inf. Rev. **42**(6), 956–970 (2018)
5. Crosato, L., Domènech, J., Liberati, C.: Predicting SME's default: are their websites informative? Econ. Lett. **204**, 109888 (2021)
6. Crosato, L., Domènech, J., Liberati, C.: Websites' data: a new asset for enhancing credit risk modeling. Ann. Oper. Res. **342**, 1671–1686 (2024)
7. Libaers, D., Hicks, D., Porter, A.L.: A taxonomy of small firm technology commercialization. Ind. Corp. Chang. **25**(3), 371–405 (2016)
8. Gök, A., Waterworth, A., Shapira, P.: Use of web mining in studying innovation. Scientometrics **102**(1), 653–671 (2015)
9. Héroux-Vaillancourt, M., Beaudry, C., Rietsch, C.: Using web content analysis to create innovation indicators-what do we really measure? Quant. Sci. Stud. **1**(4), 1601–1637 (2020)
10. Daas, P.J.H., van der Doef, S.: Detecting innovative companies via their website. Stat. J. IAOS **36**(4), 1239–1251 (2020)

11. Kinne, J., Axenbeck, J.: Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics **125**(3), 2011–2041 (2020)
12. Kinne, J., Lenz, D.: Predicting innovative firms using web mining and deep learning. PLoS ONE **16**(4), 1–18 (2021)
13. Ashouri, S., et al.: Indicators on firm level innovation activities from web scraped data. Data Brief **42**, 108246 (2022)
14. Barcaroli, G., Scannapieco, M., Donato, S.: On the use of Internet as a data source for official statistics: a strategy for identifying enterprises on the Web. Rivista Italiana di Economia Demografia e Statistica **70**(4), 25–41 (2016)
15. Bottai, C., Crosato, L., Domènech, J., Guerzoni, M., Liberati, C.: Unconventional data for policy: using big data for detecting Italian innovative SMEs. In: Proceedings of the 2022 ACM Conference on Information Technology for Social Good, Association for Computing Machinery, New York (NY, USA), pp. 338–344 (2022)

Distribution-Free Time Between Events and Amplitude Control Charts for Drought Monitoring

Michele Scagliarini(✉)

Department of Statistical Sciences, University of Bologna, Bologna, Italy
michele.scagliarini@unibo.it

Abstract. In this study, time between events and amplitude control charts are used to detect changes in the characteristics of drought events. We used non-parametric methodologies that do not require any assumption on the distribution of the phenomenon or on the observed statistics. The results indicate that the proposed methods can be valuable tools for the institutions responsible for planning drought management and mitigation measures.

Keywords: Climate Change · Control Charts · Distribution Free · Drought Index · SPEI

1 Introduction

Droughts have severe consequences for agricultural production, the environment, the economy and social stability. To properly asses this phenomenon, it is important to objectively quantify the characteristics of drought episodes in terms of their intensity, magnitude and duration. For this purpose, the Standardized Precipitation Evapotranspiration Index, SPEI, has been proposed [5]. In summary, SPEI is a climatic drought index based on precipitation, temperature and potential evapotranspiration. It can be calculated on a range of accumulation periods, from 1 to 48 months (referred to as SPEI-1, SPEI-2, and so on) and the different timescales are used to reflect the impact of drought on different water-related sectors. Meteorological and soil moisture conditions (agriculture droughts) respond to precipitation and temperature anomalies on relatively short time scales, such as 1–6 months (SPEI-1 to SPEI-6). River flow, reservoirs, and groundwater (hydrologic drought) respond to longer-term anomalies of the order of 6 months to 12 months or longer (SPEI-6 to SPEI-12). SPEI can assume positive and negative values with the following moisture categories: values greater than 2 are considered extremely wet conditions; 1.5 to 2 very wet conditions; 1 to 1.5 moderately wet conditions; -1 to 1 normal conditions; -1.5 to -1 moderately dry conditions; -2 to -1.5 severely dry conditions, and values below -2 extremely dry conditions. SPEI data are available from the Global SPEI database SPEIbase v2.9 (https://spei.csic.es/database.html) with time scales from 1 to 48 months, spatial resolution of 0.5° lat/lon (approximately 55 km), and temporal coverage from January 1901 to December 2022.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 128–133, 2025.
https://doi.org/10.1007/978-3-031-64346-0_22

The severe consequences of water scarcity conditions have focused attention on drought monitoring methods that can assist governments in implementing preparedness plans and mitigation measures to reduce the economic, environmental, and social impacts of drought. Among the possible monitoring methodologies, statistical process control methods have been found to be helpful in supporting effective drought management [4].

The relevant drought characteristics are severity, duration and frequency; therefore, a suitable statistical process control methodology for monitoring drought events should consider the time interval T between two occurrences and the magnitude X of each event. Time-Between-Events-and-Amplitude (TBEA) control charts have been proposed to monitor this type of phenomenon: a decrease in T and/or an increase in X may result in a negative condition that needs to be monitored and possibly detected with control charts. Most of the TBEA control charts proposed in the literature [2] assume known distribution functions for the variables T and X. However, in the majority of real situations, the distributions of these random variables are unknown or very difficult to identify.

In this work, we use distribution-free TBEA control charts to detect changes in the characteristics of drought events. For our purposes, aware of being non-exhaustive, we consider: the distribution-free upper sided Exponentially Weighted Moving Average (EWMA) control chart proposed by [6]; furthermore, we propose to use two non-parametric change-point control charts based on the Mann-Whitney (MW) and Kolmogorov-Smirnov (KS) statistics, respectively.

The paper is structured as follows. In Sect. 2, the methodologies considered for our purpose are introduced. Section 3 describes the data and explains and discusses the results.

2 Methodology

Let T_i ($i = 1, 2,\ldots$) be the time intervals between two consecutive occurrences of a given event E and let X_i be the corresponding magnitudes of the event of interest. To monitor T_i and X_i simultaneously [6] assumed that T_i and X_i are continuous random variables, both defined on $[0; +\infty)$, with unknown distribution functions $F_T(t|\theta_T)$ and $F_X(x|\theta_X)$ respectively, where θ_T and θ_X are known α-quantiles. Without loss of generality, it is considered that θ_T and θ_X are the median values of T_i and X_i, respectively, and when the process is in-control it follows that $\theta_T = \theta_{T_0}$ and $\theta_X = \theta_{X_0}$. The distribution-free EWMA TBEA control chart is based on the sign statistics $\mathrm{ST}_i = \mathrm{sign}\left(T_i - \theta_{T_0}\right)$ and $\mathrm{SX}_i = \mathrm{sign}\left(X_i - \theta_{X_0}\right)$, for $i = 1, 2, \ldots$, where $\mathrm{sign}(x) = -1$ if $x < 0$ and $\mathrm{sign}(x) = +1$ if $x > 0$. Let us now define the statistic $S_i = (\mathrm{SX}_i - \mathrm{ST}_i)/2$ ($i = 1, 2, \ldots$), which has the following behaviour: $S_i = -1$ when T_i increases and, at the same time, X_i decreases (positive situation); $S_i = +1$ when T_i decreases and, at the same time, X_i increases (negative situation); $S_i = 0$ when both T_i and X_i increase or when both T_i and X_i decrease (intermediate situation). Note that S_i is a discrete random variable; therefore, it would be impossible to accurately compute the run length properties of a control chart based on this statistic. To solve this problem the authors [6] developed the "continuousify" method. They propose to define an extra parameter $\sigma \in [0.1, 0.2]$ and to convert S_i into a continuous random variable, S_i^*, defined as a mixture of 3 normal random variables $Y_{i,-1} \sim N(-1, \sigma)$, $Y_{i,0} \sim N(0, \sigma)$ and $Y_{i,+1} \sim N(1, \sigma)$,

respectively. More precisely S_i^* is defined as: $S_i^* = Y_{i,-1}$ if $S_i = -1$; $S_i^* = Y_{i,0}$ if $S_i = 0$; $S_i^* = Y_{i,+1}$ if $S_i = +1$. Note that the parameter σ has to be fixed. However, the authors demonstrated that this parameter has no impact on the performance of the control chart if it falls within the suggested range. The upper-sided EWMA TBEA control chart uses the statistic $Z_i^* = \max\left(0, \lambda S_i^* + (1-\lambda) Z_{i-1}^*\right)$ with an upper asymptotic control limit given by $UCL = K\sqrt{\lambda\left(\sigma^2 + 0.5\right)/(2-\lambda)}$, where $\lambda \in [0, 1]$ and $K > 0$ are the control chart parameters and the initial value is $Z_0^* = 0$. The optimal design parameters λ and K can be obtained by studying the statistical properties of the control chart by means of a Markov chain approach [6].

As mentioned in Sect. 1, to simultaneously monitor the time T between an event E and its amplitude X, the majority of the studies proposed parametric methods. In this framework T and X are assumed to be continuous random variables, both defined on [0; $+\infty$), with known distribution functions $F_T(t|\theta_T)$ and $F_X(x|\theta_X)$ respectively, where θ_T and θ_X are the corresponding vector of parameters. Let $\mu_T = E(T)$, $\mu_X = E(X)$, $\sigma_T = SD(T)$ and $\sigma_X = SD(X)$ be the expectation and standard deviation of T and X, respectively. As usual, when the process is in control it follows that $\mu_T = \mu_{T_0}$, $\mu_X = \mu_{X_0}$, $\sigma_T = \sigma_{T_0}$ and $\sigma_X = \sigma_{X_0}$. Since the variables T and X can have very different scales, it is appropriate to define and use the normalized variables $T' = T/\mu_{T_0}$ and $X' = X/\mu_{X_0}$. Therefore, for TBEA monitoring [2] defined several dedicated statistics $Z = Z(T', X')$, functions of T' and X', satisfying the following two properties: Z increases if either T' decreases or X' increases; Z decreases if either T' increases or X' decreases. The difficulty with this approach lies in the fact that the distributions of T and X are assumed to be known. Furthermore, it is difficult to derive the distribution of the Z statistics needed to design the control chart.

Here we propose to monitor the Z statistics with non-parametric change-point control charts. In the Change Point Model (CPM) framework, it is assumed that the values of the TBEA statistic $z_1, z_2, \ldots$ Are generated by the random variables $Z_1, Z_2, \ldots$ with unknown distribution functions $F_1, F_2, \ldots$, respectively. A change point occurs at instant τ when $F_\tau \neq F_{\tau+1}$ and it is usually assumed that the observations are independent and identically distributed between every pair of change points. Therefore, the distribution of the sequence can be described by the following model: $Z_i \sim F_0$ if $0 < i \leq \tau_1$, $Z_i \sim F_1$ if $\tau_1 < i \leq \tau_2, \ldots, Z_i \sim F_k$ if $\tau_k < i \leq m$, where $0 < i < \tau_1 < \tau_2 < \cdots < \tau_k < m$ denote k unknown change points. Within this framework, it is of interest to test H_0: $k = 0$, versus H_1: $k \neq 0$, to estimate the number of change points, k, and to estimate their locations $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_k$. For testing the aforementioned hypothesis system, [1] proposed a change point control chart, designed to detect changes in the location of the process distribution, based on the Mann-Whitney (MW) U-statistic: $U_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} D_{ij}$, $1 \leq k \leq n-1$, where $D_{ij} = \text{sign}\left(Z_i - Z_j\right)$ with $D_{ij} = 1$ if $Z_i > Z_j$, $D_{ij} = 0$ if $Z_i = Z_j$, $D_{ij} = -1$ if $Z_i > Z_j$. The variance of $U_{k,n}$ depends on the split point k, so for this reason it is suitably standardized, thus obtaining the statistic $T_{k,n}$. Therefore, the test for the presence of a change point and the estimate of its time of occurrence are given by maximizing $T_{k,n}$ over k: $T_{\max,n} = \max_{1 \leq k \leq n-1} \left|T_{k,n}\right|$. To detect arbitrary changes in the process distribution, [3] proposed a control chart based on the Kolmogorov-Smirnov (KS) statistic. This

control chart tests for a change point immediately following any observation z_k by partitioning the observations into two samples, $S_1 = (z_1, \ldots, z_k)$ and $S_2 = (z_{k+1}, \ldots, z_t)$, and subsequently comparing the corresponding empirical distribution functions $\hat{F}_{S_1}(z)$ and $\hat{F}_{S_2}(z)$. The test statistic is based on the maximum difference between the empirical distributions $D_{k,t} = \sup_z \left| \hat{F}_{S_1}(z_i) - \hat{F}_{S_2}(z_i) \right|$. Further details can be found in [1] and [3].

3 Data, Results and Discussion

Among the possible timescales for the SPEI, the 12-month accumulation period was chosen because the SPEI-12 is a reliable measure of hydrological drought. The SPEI-12 data have been downloaded from the Global SPEI database for the coordinates of Bologna (44.4949° N, 11.3426° E), where observations are available from December, 1901 to December, 2022 (1453 observations).

We considered the occurrences of SPEI-12 $\leq$ -1, corresponding to drought conditions, as the events (E) of interest. Since the TBEA methodology assumes that the magnitudes of the events are defined on $[0; +\infty)$, the absolute values of SPEI-12 at the dates of occurrence have been considered as the corresponding magnitudes X_i. Given the monthly frequency of the SPEI-12, the time T_i between two events is computed in months. The available dataset contains 226 events E (values of SPEI-12 ≤ -1). The first 45 events have been used for performing the Phase I and the remaining 181 represent the monitored data. To compute the values of ST_i, SX_i, S_i and S_i^* the in-control median values have been estimated from Phase I data: $\hat{\theta}_{T_0} = 1$ and $\hat{\theta}_{X_0} = 1.2177$. Following the suggestions of the authors [6], we set $\sigma = 0.125$ and to obtain a control chart with an in-control ARL equal to $ARL_0 = 370$, we select $\lambda = 0.07$ and $K = 2.515$. Figure 1 shows the results obtained with the EWMA TBEA control chart.

As far as the change point control charts is concerned, among the possible choices for the statistics Z to be monitored, we considered the ratio $Z_R = T' \big/ X'$. To obtain the normalized variables T' and X' the in-control means have been estimated from Phase I data: $\hat{\mu}_{T_0} = 11.7333$, $\hat{\mu}_{X_0} = 1.3385$. To allow fair comparisons, the MW and KS control charts were set up with the same in-control statistical properties as the EWMA TBEA control chart ($ARL_0 = 370$). The results are shown in Fig. 2, which reports the estimated change points, the values of the Z_R statistic, and the averages of the observed statistics between each pair of change points (dashed line).

By examining the results, it can be seen that for Phase I data no alarms are reported; therefore, this dataset is in statistical-control and the estimated medians $\hat{\theta}_{T_0} = 1$, $\hat{\theta}_{X_0} = 1.2177$ and averages $\hat{\mu}_{T_0} = 11.7333$, $\hat{\mu}_{X_0} = 1.3385$ can be considered as reliable estimates of the corresponding unknown parameters and used for the monitoring phase. Before commenting further on the results, it should be remembered that we have used different methodologies and therefore we do not expect identical results. However, it can be noted that the control charts considered agree in establishing that the drought evolved for the worse during our study period: since August 2003 (MW and KS control charts, Fig. 2) or August 2012 (EWMA control chart, Fig. 1), the TBEA statistics show significant increases. Furthermore, it is worth noting that the examined methodologies provide results that can be used in a complementary way. As far as the EWMA chart is concerned, it is possible to perform a post-signal analysis to highlight out-of-control

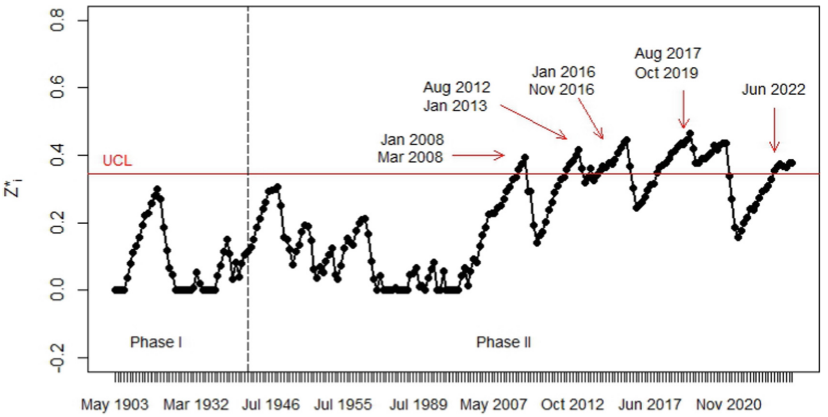


Fig. 1. Time Between Event and Amplitude EWMA control chart

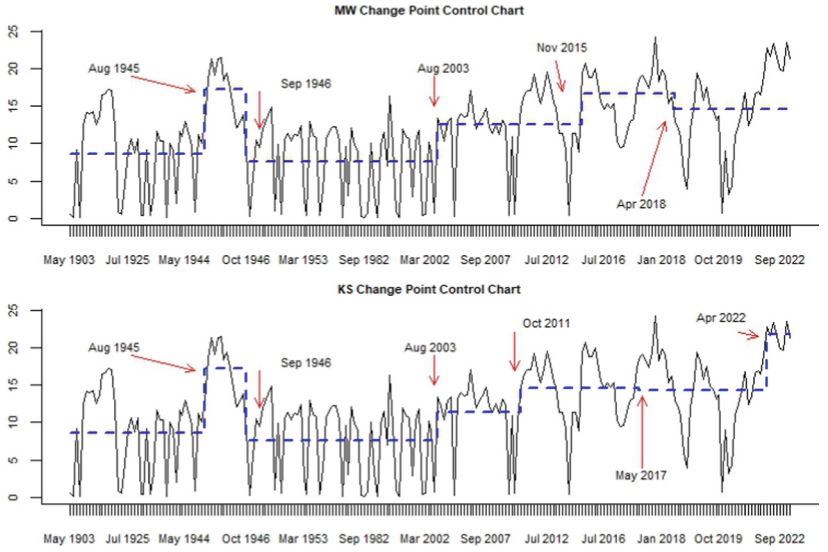


Fig. 2. Mann-Whitney (upper panel) and Kolmogorov-Smirnov (lower panel) control charts

events that are consecutive with respect to the original time scale. For example, note the long period of temporally consecutive out-of-control Z_i^* values between August 2017 and October 2019. To discuss the results obtained with the change point control charts, it is worth remembering that we are monitoring $Z_R = T' \big/ X'$, which simultaneously takes into account the time between drought events and their magnitudes. This means that by detecting change points in the sequence of this TBEA statistic, it is possible to identify significant changes in drought characteristics. For example, consider the MW control chart. The observed increases in the location of the TBEA statistic in August 2003 and November 2015 indicate an increase in drought severity.

In summary, since drought mitigation measures vary according to drought severity and duration, TBEA control charts can provide valuable input to institutions responsible for planning suitable drought management and designing resilience policies.

References

1. Hawkins, D.M., Deng, Q.: A nonparametric change-point control chart. J. Qual. Technol. **42**(2), 165–173 (2010)
2. Rahali, D., Castagliola, P., Taleb, H., Khoo, M.B.C.: Evaluation of shewhart time-between-events-and-amplitude control charts for several distributions. Qual. Eng. **31**(2), 240–254 (2019)
3. Ross, G.J., Adams, N.M.: Two nonparametric control charts for detecting arbitrary distribution changes. J. Qual. Technol. **44**(12), 102–116 (2012)
4. Townsley, J., Chimka, J.R.: A control chart for severity index to detect drought compare favourably to logistic regression. Hydrol. Res. **42**(1), 1–9 (2011)
5. Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I.: A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J. Clim. **23**(7), 1696–1718 (2010)
6. Wu, S., Castagliola, P., Celano, G.: A distribution-free EWMA control chart for monitoring time-between-events-and-amplitude data. J. Appl. Stat. **48**(3), 434–454 (2021)

Small Area Estimation of Educational Poverty Using Item Response Theory Models

Maria Giovanna Ranalli[1(✉)] and Gaia Bertarelli[2]

[1] Department of Political Science, University of Perugia, Perugia, Italy
maria.ranalli@unipg.it
[2] Department of Economics, Ca' Foscari University of Venice, Venice, Italy

Abstract. Educational Poverty (EP) is a concept of increasing importance whose measure, particularly at local level, is of paramount relevance to enforce and monitor policies. Measuring EP is hindered by its latent and possibly multidimensional nature, and researchers have not yet agreed on a set of items that would serve this aim. In this paper, we focus on a set of 33 binary indicators measured with the Activities of Daily Living survey on a sample of 4,382 individuals aged 15–29, and propose to use multidimensional Item Response Theory (IRT) models to extract the latent dimensions of EP. To obtain estimates of these latent dimensions at local (sub-regional) level, we embed Small Area Estimation in the multidimensional IRT model by allowing the values of the latent factors to change with covariates and a area-specific random effects.

Keywords: Latent regression model · random effects · unit-level model · binary items

1 Background

In recent years, Educational Poverty (EP) has emerged as a concept of increasing importance in scientific literature on social phenomena. The relevance of EP in well-being and sustainable development is undoubtedly recognised in the wider debate on the "Agenda 2030 for Sustainable Development Goals" which highlights the need to "provide everybody with a quality, equal and inclusive education and permanent learning opportunities".

The study of EP in Italy dates back to a first tentative definition provided by Save the Children [1], where EP is defined as "the impossibility for children and teenagers to learn, experiment, develop and freely foster their capacities, talents and aspirations". Then, Quattrociocchi [2] introduces an EP index for the youngsters (individuals aged 15–29) using a set of 21 indicators, grouped into four dimensions: (*i*) Participation, measuring the lack of participation of youngsters to the social life; (*ii*) Resilience, measuring the lack of development of an attitude of trusting oneself and one's abilities; (*iii*) Standard of living, measuring the lack of the opportunity to lead an inclusive, healthy, and safe life

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 134–139, 2025.
https://doi.org/10.1007/978-3-031-64346-0_24

having an adequate standard of living; (*iv*) Friends and skills, representing the lack of the opportunity to wave relationships with others and to achieve those skills (digital, literacy and numerical skills) needed to succeed in such a fast-pacing world. The 21 indicators used to measure these four dimensions come from several data sources: (*i*) the Aspects of Everyday Life survey (Aspetti della Vita Quotidiana - AVQ), (*ii*) the EU Statistics on Income and Living Conditions, (*iii*) the European Health survey, (*iv*) the Labor Force survey, (*v*) the Time Use survey and (*vi*) the International Assessment of Adult Competencies. These 21 indicators include both simple indicators resulting from specific questions or composite indicators resulting from the aggregation of multiple items from the same survey. Quattrociocchi [2] measures EP at macro-regional level (North-east, North-west, Center, South).

Pratesi et al. [3] recognize the importance of obtaining EP estimates at sub-regional level, as EP measurements can be crucial to tailor policy actions at local level, and use Small Area Estimation (SAE) techniques to obtain a composite indicator of EP for 59 small areas obtained by the intersection of the Italian Regions (NUTS 2 level) with the degree of urbanization as measured by Eurostat with the DEGURBA three-level classification: cities, towns and suburbs, urban areas. They propose an area-level two-step approach: in the first step they apply an area-level SAE model for each indicator to obtain more reliable estimates, which are then aggregated in the second step following the approach proposed by Mazziotta and Pareto [4]. In order to obtain data at the desired level, Pratesi et al. use only 14 indicators (simple items or composites) from the AVQ survey grouped into four dimensions. Bertarelli et al. [5] provide a fuzzy multidimensional measure of EP at regional level using an index based on 53 different items available from the AVQ survey.

In our view, the definition of an EP index can be enhanced by viewing it as a multidimensional latent construct hidden behind a series of observable binary indicators, and we propose to extract these latent dimensions using multidimensional Item Response Theory (IRT) models. See, e.g., Skrondal and Rabe-Hesketh [6] for an introduction to IRT models. Then, existing SAE methods must be adapted to obtain estimates of EP and of its dimensions at sub-regional level. Latent variable models have been considered in a SAE framework in Fabrizi et al. [7], where a latent class unit level regression model is introduced for extracting and estimating disability small area counts from survey data. Giovinazzi and Cocchi [8] take a similar approach to measure social integration of second generation students in the Italian school system. Moretti et al. [9] use factor analysis models to extract economic well-being and apply SAE unit-level models using the factor scores as outcome variables.

In this paper, we take a one-step perspective and propose to estimate EP at local level using a multidimensional mixed effects IRT model [10], which includes covariates to obtain predictions. Specifically, we use 33 AVQ items and focus on the 59 small areas considered in [3]. This paper is a methodological extension of [11] where a two-step approach is used to deal with this estimation problem and is organized as follows. Section 2 describes the available data in detail, while Sect. 3 illustrates the proposed modeling approach. Section 4 provides some preliminary conclusions.

2 Data

In this work, following [3,5], we use only items coming from the AVQ survey 2016. In particular we focus on 33 items from the 2016 edition with no missing values. See Table 1.

Table 1. Items used from AVQ and corresponding description.

Item	Description
VOLON	0 if one has volunteered by free choice now or in the past, 1 otherwise
VOL POL	0 if one hasvolunteered for a political party in the last year; 1 otherwise
POL PROC	0 if one hasparticipated in political/social demonstrations in the last year; 1 otherwise
MUSIC	0 if one goes to music concerts at least 4 times a year, 1 otherwise
MUSEUM	0 if one goes to museum at least 4 times a year, 1 otherwise
CINE	0 if one goes to cinemas at least 4 times a year, 1 otherwise
THEATER	0 if one goes to theatres at least 4 times a year, 1 otherwise
SPORT	0 if one does sport continuously, 1 otherwise
BOOK	0 if one reads/has read books without obligation, 1 otherwise
DIRTY	0 if one lives in an area considered clean, 1 otherwise
PARKS	0 if one lives in an area with parks, 1 otherwise
CRIME	0 if one lives in a neighborhood not at risk of crime, 1 otherwise
POLLUT	0 if one lives in an area with significant pollution problems, 1 otherwise
LIGHT	0 if one lives in an area with lighted streets, 1 otherwise
INT INF	0 if one uses the internet several times a week, 1 otherwise
PCOPE2A	0 if one handles files in an operating system, 1 otherwise
CLOUD	0 if one is capable of organizing his work on the cloud, 1 otherwise
INTCOMU5	0 if someone has sent at least one email in the last 3 months, 1 otherwise
INTCOMU6	0 if one has interacted at least once on social networks in the last 3 months, 1 otherwise
INTCOMU1B	0 if one has made at least one video call in the last, 1 otherwise
PCOPE10A	0 if one has exchanged files between different devices in the last three months, 1 otherwise
PCOPESW	0 if one knows how to install software and applications, 1 otherwise
PCOPEIM	0 if one knows how to change software settings, 1 otherwise
PCOPEWO	0 if one knows how to use Word, 1 otherwise
PCOPEX	0 if one knows how to use Excel, 1 otherwise
PCOPEXC	0 if one uses advanced Excel options, 1 otherwise
PCOPESL	0 if one knows how to make presentations with graphs, figures and tables, 1 otherwise
INTUSO1	0 if one has independently used the internet to interact with the public administration in the last 3 months, 1 otherwise
FRIEND	0 if one meets friends at least 4 times a month, 1 otherwise
FRIEND2	0 if one has at least one friend to count on, 1 otherwise
NEIGH	0 if one has at least one neighbor to count on, 1 otherwise
TRUST	0 if one claims to trust people, 1 otherwise

The data set used in the empirical analysis covers 4,382 individuals aged 15–29. The population size is equal to 9,178,438. The sample size in the 59 considered domains goes from a minimum of 11 individuals, to a maximum of 295. Half of the domains considered in our analysis have a maximum sample size of 62.

Covariates play a role of primary importance in SAE. Here, we have individual-level covariates, such as gender, employment condition, nationality, and municipality-level covariates measuring social, cultural, and IT assets. In the following section we introduce notation and the proposed modelling approach.

3 The Estimation Problem and the Proposed Approach

Let U denote the finite population of size N, which can be partitioned into m domains (small areas) U_i with size N_i, for $i = 1, \ldots, m$, $m = 59$. Let $y_{ij\ell}$ denote the value of binary item ℓ on unit j in area i, $\ell = 1, \ldots, L$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. Let s_i denote the set containing the n_i population indexes of the sample units belonging to area i, with $n = \sum_{i=1}^{m} n_i$. Now, suppose there are K latent factors that represent K different dimensions of EP, then $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ijK})^T$ is their value on unit j in area i. The estimation problem in this paper is to first identify the number of dimensions K and then quantify population means of these factors, that is

$$\bar{\theta}_{ik} = \frac{1}{N_i} \sum_{j=1}^{N_i} \theta_{ijk}, \quad \text{for } k = 1, \ldots, K \text{and} i = 1, \ldots, m. \tag{1}$$

Therefore, this estimation problem is twofold: on one side, it requires latent variable modelling to identify the latent traits from the observed items, then it needs SAE methods to address the limited information, in terms of sample units, coming from the areas. A multidimensional IRT model can be used to extract the latent traits from the set of $L = 33$ AVQ items. For example, in a two-parameter logistic (2PL, [12]) model the probability for unit j in area i of scoring 1 for item ℓ depends on the latent traits, $q_{ij\ell} = P(y_{ij\ell} = 1 | \theta_{ij1}, \ldots, \theta_{ijK})$, and is modeled as

$$\text{logit}(q_{ij\ell}) = \alpha_{\ell 0} + \sum_{k=1}^{K} \alpha_{\ell k} \theta_{ijk}, \tag{2}$$

where $\alpha_{\ell 0}$ is a difficulty parameter for item ℓ, while $\alpha_{\ell k}$ measures the discrimination power of item ℓ for latent trait k, for $k = 1, \ldots, K$ and $\ell = 1, \ldots, L$. Model (2) would essentially be a logistic regression model, if it weren't for the fact that the covariates θ_{ijk}'s are not observed. Maximum likelihood can be used to obtain parameter estimates for $\alpha_{\ell 0}$ and $\alpha_{\ell k}$, $k = 1, \ldots, K$, under the assumption that item responses are independent given the latent variables (conditional independence assumption), and that θ_{ijk}'s are independent standard Gaussian random variables.

Values of the latent traits can be assigned to all sampled units using posterior predictions for θ_{ijk}'s given the data. If sample sizes were large enough, these predicted traits could be used to compute direct estimates for area means in (1). Instead, it is required to borrow strength using information from other areas through a model, where auxiliary information plays a central role. Let $\boldsymbol{x}_{ij}$ denote

a p dimensional vector of auxiliary variables for unit j in area i, for which we know population level area means $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}/N_i$.

In the literature on SAE for latent outcomes, this is usually accomplished in two steps: first the latent variables are estimated and then used as (if they were observed) response variables in a SAE model. However, this approach may yield biased parameter estimates for the relationships between the latent variable and the covariates so that errors are propagated from the first step to the second without any control. This makes it very challenging to obtain a reliable measure of uncertainty of the final small area estimates. One noticeable exception is in [7] where the problem of classifying the population and obtaining small area estimates is tackled in one step within a hierarchical Bayesian framework in which the probability of belonging to each latent class changes with covariates. In this paper, we also take a one-step perspective for the case of multiple continuous outcomes, and embed SAE in a multidimensional IRT model in which the values of the latent factors change with covariates. In particular, we assume that the following model holds for all j and i:

$$\text{logit}(q_{ij\ell}) = \alpha_{\ell 0} + \sum_{k=1}^{K} \alpha_{\ell k}\theta_{ijk} \quad \text{for} \ell = 1, \ldots, L, \tag{3}$$

$$\theta_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + v_{ik}, \tag{4}$$

where $\boldsymbol{\beta}_k$ is a vector of fixed coefficients for the regression of the covariates on latent construct θ_k, v_{ik} is an area-level Gaussian random effect with mean zero and variance σ^2_{vk}, and $\theta_{ij1}, \ldots, \theta_{ijK} \sim MVN(\mathbf{0}, \boldsymbol{I}_K)$. Numerical methods, such as the Metropolis-Hastings Robbins-Monro stochastic imputation algorithm can be used to maximize the likelihood and to estimate fixed and random coefficients (see [10]). Once these are obtained, a model-based estimate of (1) can be computed in the spirit of the unit-level model by Battese et al. [13] as

$$\hat{\bar{\theta}}_{ik} = \bar{\boldsymbol{x}}_{i^T} \hat{\beta}_k + \hat{v}_{ik}, \quad \text{for } k = 1, \ldots, K. \tag{5}$$

The choice of the number of latent dimensions K is made using information criteria. In our application, we find five latent dimensions with an interesting interpretation. The variance of the final small area estimates in (5) is obtained using bootstrap.

4 Preliminary Remarks

IRT models provide a particularly suitable methodological framework in research contexts where latent and multidimensional constructs are the objects of analysis. To the best of our knowledge, in this short paper we propose a first attempt to integrate IRT models in the SAE framework to study EP for younger people (15–29) at sub-regional level. The proposed approach can be generalized to other possible data or set of items as long as they are available at the unit level. This makes our method an appropriate model for possible future analyses on EP, a phenomenon which is still under study from a definitional point of view.

Some methodological limitations will need to be addressed as future work. Among these, the treatment of missing data is of primary importance in unit-level models that consider many items as manifest variables. Furthermore, some reflections on the method of aggregating several latent variables into a single measure need to be conducted. Indeed, in this work, once the latent dimensions have been identified, a final overall estimate for EP at local level is provided using a bi-factor model. Other models, such as the higher-order model, from the confirmatory analysis toolbox can be used as well.

References

1. Save The Children. (2015). Illuminiamo il futuro 2030—Obiettivi per liberare i bambini dalla Povertã Educativa. Accessed 02 March 2024. https://www.savethechildren.it/cosa-facciamo/pubblicazioni/illuminiamo-il-futuro-2030-obiettivi-liberare-i-bambini-dalla-povert%C3%A0.
2. Quattrociocchi, L.: Povertã educativa: (Non) finirai come tuo padre.Slide istat, popolazione, istruzione, mobilitã : Giornata di studio in collaborazione tra AISP, SIS, Istat e SIEDS (2018). https://www.slideshare.net/
3. Pratesi, M., Quattrociocchi, L., Bertarelli, G., Gemignani, A., Giusti, C.: Spatial distribution of multidimensional educational poverty in Italy using small area estimation. Soc. Indic. Res. **156**, 563–586 (2021)
4. Mazziotta, M., Pareto, A.: Measuring well-being over time: the adjusted Mazziotta-Pareto index versus other non-compensatory indices. Soc. Indic. Res. **136**, 967–976 (2018)
5. Bertarelli, G., D'Agostino, A., Giusti, C., Pratesi, M.: measuring educational poverty in Italy: a multi-dimensional and fuzzy approach. In: Analysis of Socio-Economic Conditions: Insights from a Fuzzy Multidimensional Approach, pp. 166–179. Routledge (2021)
6. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. CRC Press (2004)
7. Fabrizi, E., Montanari, G.E., Ranalli, M.G.: A hierarchical latent class model for predicting disability small area counts from survey data. J. R. Stat. Soc. Ser. A **179**, 103–132 (2016)
8. Giovinazzi, F., Cocchi, D.: Social integration of second generation students in the Italian school system. Soc. Indic. Res. **160**, 287–307 (2022)
9. Moretti, A., Shlomo, N., Sakshaug, J.W.: Small area estimation of latent economic well-being. Sociol. Methods Res. **50**(4), 1660–1693 (2021)
10. Chalmers, R.P.: Extended mixed-effects item response models with the MH-RM algorithm. J. Educ. Meas. **52**, 200–222 (2015)
11. Bertarelli, G., Ranalli, M.G., Pratesi, M.: Multivariate small area estimation of educational poverty with latent variable models. In: 15th International Conference of the ERCIM Working Group on Computational and Methodological Statistics, London, 17–19 December 2022, Book of Abstracts, 75 (2022)
12. Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (eds.) Statistical Theories of Mental Test Scores, pp. 395–479. Addison-Wesley, Reading, MA (1968)
13. Battese, G.E., Harter, R.M., Fuller, W.A.: An error-components model for prediction of county crop areas using survey and satellite data. J. Am. Stat. Assoc. **83**(401), 28–36 (1988)

Estimating Multidimensional Educational Poverty in Italy Using a Quantile Approach

Caterina Giusti(✉) and Francesco Schirripa Spagnolo

Department of Economics and Management, University of Pisa, Pisa, Italy
{caterina.giusti,francesco.schirripa}@unipi.it

Abstract. In recent years, addressing Educational Poverty (EP) has emerged as a pressing issue on the political agendas of several countries, recognized as a new social challenge requiring urgent attention. To assess EP in Italy, the Italian National Statistical Institute introduced a multidimensional composite index known as the EPI. Considering the same dimensions of this index and the relationships among them, we employ a quantile approach to have a comprehensive understanding of the relationships among the variables. Our results suggest that the EPI dimensions play a different role at different points of the index distribution.

Keywords: Quantile composite-based path modelling · Composite indicators · Educational inequalities

1 Introduction

Recently, Educational Poverty (EP) has become increasingly important in several countries. EP is usually considered a multidimensional phenomenon, and different definitions and measures are currently used to define it. In Italy, a multidimensional definition of EP has been proposed by [7] which defines EP as the "*deprivation by children and adolescents of the possibility of learning, experimenting, and freely developing skills, talents and aspirations*". Moreover, Save the Children also developed an educational poverty index composed by nine indicators that measure the access to educational services from early childhood, the quality of the education offered at school, but also participation in sports, cultural, and recreational activities, school dropout, and levels of skills achieved [7]. This attempt to formulate a multidimensional index for educational poverty was followed by ISTAT, the Italian National Statistical Institute, which proposed a composite index, called EPI, measuring a combination of material, relational, cultural and environmental problems that can limit the ability of aged 15–29 to live in the society [6]. The dimensions of this index are: '*Participation*', '*Resilience*', '*Standard of living*', and '*Friends & skills*'.

This work aims at extending the analysis by [6] measuring EP in Italian regions while also considering gender as an additional key factor. Indeed, educational gender inequalities are present in developed European countries [4]. At

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 140–145, 2025.
https://doi.org/10.1007/978-3-031-64346-0_24

the same time, Italy is historically characterized by a North-South divide for many economic and social phenomena. Therefore, disaggregating the EPI for gender and regional level allows us to control for these effects and to provide information that can be used to drive local policies that are often managed at the regional level.

Moreover, in this work we study the EPI applying the quantile composite-based path modelling (QC-PM), a quantile approach in the partial least squares path modelling (PLS-PM) framework [2]. This approach allows us to evaluate whether and how much the impact of each domain of EPI varies among units with a high, medium or low level of this index.

2 Educational Poverty Index

To estimate the EPI we only use the items from the multipurpose survey on households "Aspects of daily life" (AVQ), carried out by the Italian National Statistical Institute (ISTAT) each year to collect information on citizens' habits and problems they face in everyday life [3]. Specifically, our data are from the 2021 AVQ wave[1]. As the target population, we select individuals aged 14–34 years and our estimation sample contains 6827 observations with nonmissing variable information. According to the definition of ISTAT [6] and following the works of [1,5], we consider four dimensions to measure the EP: *Participation*, *Resilience*, *Standard of living* and *Friends & skills*. Each dimension is measured by items/indicators obtained from the AVQ.

The first domain of the EPI is *Participation* (X1) and it aggregates three indicators measuring the participation of youth in social life:

- Political participation (X1.1): the percentage of people who do not carry out political activity;
- Internet usage (X1.2): the percentage of people who do not use the internet or do not use it every day;
- Volunteering (X1.3): the percentage of people who do not volunteer freely.

The second domain (X2), *Resilience*, represents the development of an attitude towards trusting oneself and one's abilities, and it is composed by six indicators:

- Free-time satisfaction (X2.1): the percentage of people who are not satisfied with their free time;
- Rely on network (X2.2): the percentage of people who do not have people to rely on;
- Cultural activity (X2.3): the percentage of people who perform fewer than 4 cultural activities;
- Book reading (X2.4): the percentage of people who do not read books or read fewer than four books a year;

[1] The data are freely available on the ISTAT website: https://www.istat.it/en/archivio/129959.

– Trusting (X2.5): the percentage of people who do not trust others;
– Public administration interaction (X2.6): the percentage of people who do not use the internet to interact with PA.

Three indicators are used to measure the third domain (X3), *Standard of living*, which represents the ability to live an inclusive, healthy, and safe life while providing an appropriate quality of living:

– Sport practice (X3.1): the percentage of people who do not practice sports;
– Deterioration in one's surroundings (X3.2): the percentage of people who live in the presence of elements of deterioration in the areas in which they live;
– Green spaces (X3.3): the percentage of people who do not have public parks where they live.

The fourth and last domain (X4), *Friends & skills*, represents the capacity to build relationships with others and to develop the abilities required to flourish in a fast-paced environment is composed by two indicators:

– Friend network (X4.1): the percentage of people who do not have friends or who rarely have friends;
– Digital skills (X4.2): the percentage of people with low digital skills, represented by a combination of software skills, problem-solving ability, communication, and information skills.

3 Methodology

We use a hierarchical composite model for measuring and modelling the EPI in Italian regions by gender considering the EPI as the highest-order composite indicator, and the other four domains as constructs of lower order. The path diagram in Fig. 1 shows the specified hierarchical structure of the EPI and how the constructs are connected.

To study the conditional distribution at locations different from the mean we decided to use the quantile composite-based path modelling (QC-PM) [2]. The algorithm of the QC-PM approach uses Quantile Regression (QR) instead of OLS regression to estimate the model parameters. In particular, for each quantile of interest, $\tau \in (0, 1)$, in the first stage through an iterative algorithm alternating simple QR or multiple QR, the outer weights $w_{jk}(\tau)$ and each latent construct ξ_j are estimated. In the second stage, the loading coefficients and the path coefficients are estimated through QR:

$$\xi_j(\tau) = \lambda_{0j}(\tau) + \boldsymbol{\lambda}_{jk}(\tau)\mathbf{x}_{jk} + \varepsilon_j, \tag{1}$$

$$\xi_j(\tau) = \beta_0(\tau) + \sum_i \boldsymbol{\beta}_{ji}(\tau)\xi_i + e_j, \tag{2}$$

Therefore, this approach provides a set of outer weights, loadings and path coefficients for each quantile of interest. Inference in the model is performed via a bootstrap technique.

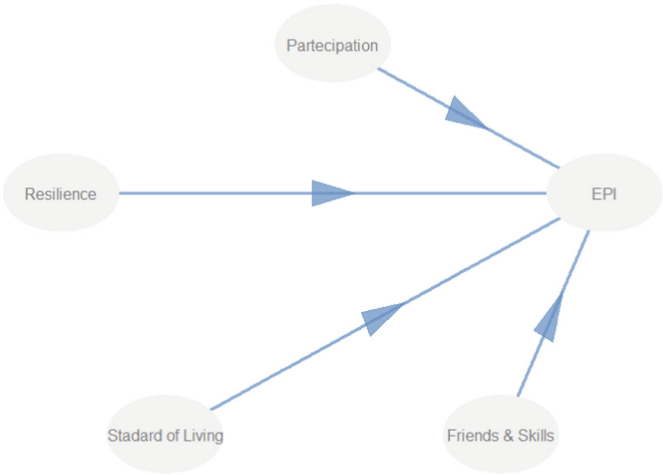


Fig. 1. Path diagram.

4 Results and Discussion

We fitted the model to five selected quantiles, namely, 0.10, 0.25, 0.50, 0.75, and 0.90. Table 1 shows the estimation of the impact of each domain on the EPI (namely, the path coefficients).

Table 1. Path coefficients†

	Results for the following value of q:				
Domain	0.1	0.25	0.5	0.75	0.9
Participation	0.317^{***}	0.274^{***}	0.292^{***}	0.271^{***}	0.277^{***}
Resilience	0.317^{***}	0.336^{***}	0.340^{***}	0.315^{***}	0.196^{***}
Standard of Living	0.300^{***}	0.270^{***}	0.254^{***}	0.250^{***}	0.283^{***}
Friends & Skills	0.316^{***}	0.333^{***}	0.282^{***}	0.333^{***}	0.426^{***}

†p-value: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Each domain has a positive and significant effect on the EPI at each of the five selected quantiles and this impact differs across the distribution of the index. The effect associated with the *Participation* domain is greater at the bottom of the distribution than at the other parts. This might indicate that engagement in social activities is more important for units (intersection between region and gender) experiencing lower levels of educational poverty. The path coefficient of the *Resilience* domain is constant for the first four quantiles and then decreases at $q = 0.90$: this indicates that the attitude towards trusting oneself and one's abilities is less important for units with a higher level of EP than for those

with a lower level of EP. The path coefficient associated with the *Standard of Living* dimension follows a U-curve that is greater at the extreme of the EPI distribution than at the centre. This highlights that the effect of the ability to lead an inclusive, healthy, and safe life is the same at the lowest and the highest levels of EP. The effect for the last dimension, *Friends & Skills*, is greater at the top of the distribution than at the bottom. Therefore, these results suggest that the effect of the four considered dimensions on the EPI depends on the level of EP by region and gender.

In Fig. 2, we plot the EPIs obtained by fitting the QC-PM at $q = 0.10$, 0.50 and 0.90 for females and males in the Italian regions. First of all, we can observe the well-known North-South divide, with higher estimates of EP affecting southern regions at all quantiles. Then, in general, moving from lower to higher quantiles, the level of EP decreases for females, while the opposite is observed for males. For instance, females in Tuscany (Central Italy) have low levels of education poverty at all considered quantiles, while males tend to have higher

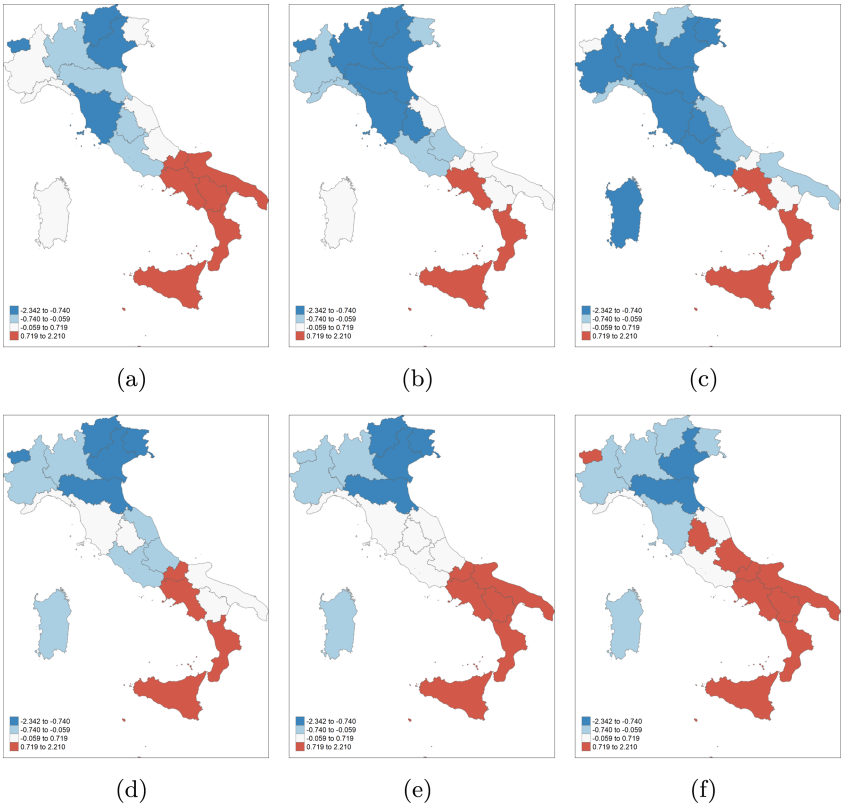


Fig. 2. Map of the EPI by region obtained by fitting the QC-PM at for females at $q = 0.10$ (a) $q = 0.50$ (b), $q = 0.90$ (c) and males at $q = 0.10$ (d) $q = 0.50$ (e), $q = 0.90$ (f)

levels, especially at $q = 0.10$ and $q = 0.50$. In contrast, males in Sardinia tend to have lower levels of EPI at $q = 0.10$ and $q = 0.50$, while for the higher quantile the situation is the opposite. Further, in Piemonte (northwest), for the lowest level of the EPI, females are more deprived; however, when we look at the upper part of the EPI distribution, males seem to be more deprived than females. A similar pattern is observed also for the Puglia region in southern Italy.

5 Final Remarks

The results obtained by employing a quantile composite approach to the EPI showed that the four dimensions play different roles when considering the full distribution of the index. In particular, controlling for regional and gender differences, *Standard of living* and *Friends & Skills* are the most important dimensions at higher levels of EP. Regarding the geographical disparities, as expected, the Italian southern region are the most deprived. Considering the differences by gender, the level of EP of females in several regions of northern and central Italy seems to decrease when moving from low to higher quantiles. In contrast, focusing on the male population, the regions characterized by high levels of EP are more numerous when considering the higher quantiles, not only in the southern regions but also in some northern regions.

References

1. Bertarelli, G., D'Agostino, A., Giusti, C., Pratesi, M.: measuring educational poverty in Italy: a multi-dimensional and fuzzy approach In: Betti, G., Lemmi, A. (ed.) Analysis of Socio-Economic Conditions: Insights from a Fuzzy, pp. 166–179 (2021)
2. Davino, C., Vinzi, V.E.: Quantile composite-based path modeling. Adv. Data Anal. Classif. **10**(4), 491–520 (2016). https://doi.org/10.1007/s11634-015-0231-9
3. Istat: Aspects of daily life survey, Istat, 2023
4. OECD: The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence, OECD Publishing (2015)
5. Pratesi, M., Quattrociocchi, L., Bertarelli, G., Gemignani, A., Giusti, C.: Spatial distribution of multidimensional educational poverty in Italy using small area estimation. Soc. Indic. Res. **156**, 563–586 (2021)
6. Quattrociocchi, L: Povertà educativa:(Non) finirai come tuo padre,slide istat , popolazione, istruzione, mobilità: Giornata di studio in collaborazione tra AISP, SIS, Istat e SIEDS.https://www.slideshare.net/slideistat (2018)
7. Save the Children: La lampada di Aladino. https://s3.savethechildren.it/public/files/uploads/pubblicazioni/la-lampada-di-aladino.pdf (2014)

New Challenges for Measuring Multidimensional Educational Poverty in Official Statistics

Elisabetta Segre(✉), Miria Savioli(✉), and Valeria Quondamstefano(✉)

ISTAT, Italian National Institute of Statistics, Rome, Italy
{elisabetta.segre,miria.savioli,valeria.quondamstefano}@istat.it

Abstract. This paper outlines the work of a Scientific Commission led by Istat, which involved multiple institutions working together to assess and map educational poverty in Italy. The Commission has developed a theoretical framework that identifies two critical aspects of educational poverty that must be addressed simultaneously: poverty in outcomes and poverty in drivers (resources). The Commission's ultimate goal is to validate a dashboard of multisource indicators, uncover data gaps, and define the most appropriate methodology to synthesize the information. To support the Commission in finalizing its work, we carried out an exploratory exercise synthesizing and mapping indicators selected from a preliminary list of indicators defined by the Commission. We calculated two composite indicators, one for the outcomes and one for the drivers. This process has provided valuable insights into the challenges the Commission will face going forward.

Keywords: educational poverty · multidimensional framework · composite indicator

1 Introduction

In 2023, Istat led an Inter-Institutional Scientific Commission to define and measure educational poverty as a multidimensional phenomenon to inform public policies to fight educational poverty and guarantee equal opportunity. The Commission is made up of over 50 members, including representatives from academia, UNICEF, World Bank, UNESCO, Save the Children, Ministry of Education, Bank of Italy, Guarantor Authority for Children and Adolescents, Fondazione Con i Bambini, ANCI, and INPS. They are expected to finish their work by the end of 2024.

After analyzing the relevant international and national literature [1–7, 14, 15] and existing sources of information, the Commission defined a conceptual framework for measuring the phenomenon and identified an initial broad set of multi-source indicators that can be calculated from the databases available within the national statistical system or, in the future, from new sources or ad hoc surveys.

By the end of 2024, the Commission should arrive at a final set of indicators and a comprehensive list of unmet information needs. In addition, the Commission will explore methods and processes to synthesize the information using composite indicators and to

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 146–151, 2025.
https://doi.org/10.1007/978-3-031-64346-0_25

provide a mapping of the territory, allowing perimeter priority areas towards which direct investments and interventions.

In this brief document, we explain the theoretical framework established by the Commission based on relevant national and international literature (Sect. 2). Additionally, we present the methods (Sect. 3) and results (Sect. 4) of a quantitative analysis performed to assess the framework's efficacy and to provide the Commission with valuable insights into the challenges that lie ahead (Sect. 5).

2 Conceptual Framework for Educational Poverty

The Commission decided to focus on the period of life when most human and social capital is built up: childhood and adolescence (0 to 19 years old), and, inspired by the OECD's framework for measuring child well-being [11], identifies two dimensions of educational poverty: outcome poverty and drivers (resources) poverty.

Poverty of resources refers to a condition that arises from a lack of educational and cultural resources available to the community in its broadest sense, including family, schools, and places of learning and engagement. It can also result from limited access to valuable experiences for personal growth.

On the other hand, poverty of outcomes concerns the acquisition of cognitive and non-cognitive skills that enable individuals to grow and develop relationships with others, cultivate their talents, and realize their aspirations, together with a sense of being part of a community, the ability to consciously exercise the right of active citizenship and contribute positively to the well-being of the community.

Due to the phenomenon's complexity, each of the dimensions considered needs to be broken down into sub-dimensions to be measured effectively (Fig. 1). Two subdimensions were identified for outcome poverty: cognitive and non-cognitive skills (emotional, relational, trusting interactions), while three subdimensions were identified for resource poverty: family, school, and a broader social and cultural context in which children live.

A broad set of indicators has been proposed for each sub-dimension to capture its main components. In the drivers' dimension, both resource/opportunity indicators and accessibility/participation indicators were selected. Beyond this choice, the idea is to monitor the presence of the facility/service (school, library, theater, and so on) and the actual habit of accessing it. In the dimension on outcomes, INVALSI's achievement tests, administered to all students in second, fifth, eighth, tenth, and thirteenth grade, monitor cognitive skills well. On the other hand, assessing individual non-cognitive skills such as creativity, curiosity, self-esteem, motivation, adaptability, stress management, cooperation, and communication remains a challenge, and official statistics must address this data gap.

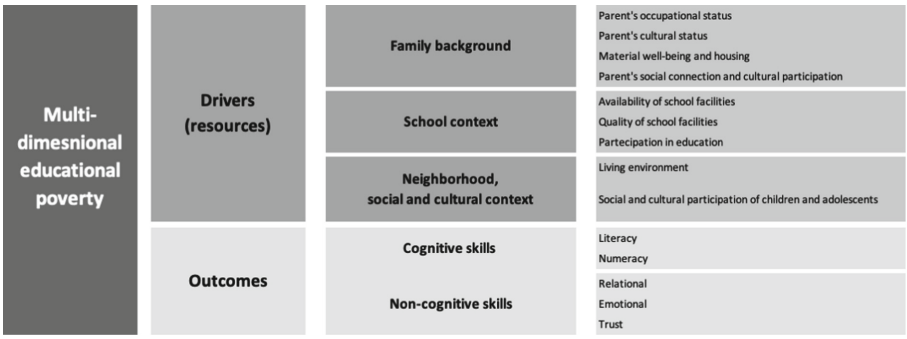


Fig. 1. Commission's theoretical framework for measuring educational poverty of children and adolescents (0–19 years old).

3 Methods and Data

To support the Commission finalizing its work, we conducted an exploratory exercise to assess territorial gaps in deprivation in outcomes and drivers (resources). We selected a set of indicators from a preliminary list defined by the Commission, and we calculated two composite indicators, one for the outcomes and one for the drivers. This process has provided valuable insights into the challenges the Commission will face going forward.

The indicators used in this exercise are a subset selected among those proposed by the Commission calculated on data available at the municipality level. Despite aiming to map the phenomena at a small territorial scale, municipalities in Italy are not the most suited geographical areas because of the great number of very small cities in terms of land and population. For this reason, each indicator is computed at a geographical level that combines administrative regions and municipalities' degrees of urbanization, which are classified as cities, small towns or suburbs, and rural areas (known as Degurba classification). This means that for each indicator and within each region, all municipalities with the same degree of urbanization report the same value for that indicator.[1]

To partially compensate for the lack of information on noncognitive skills, we included in the outcome dimension four indicators to monitor the difficulties encountered in completing the upper secondary cycle, which can be traced back to a plurality of factors, including noncognitive ones.

The indicators are meant to identify possible scenarios of risk or deprivation[2], as commonly recognized in the international literature on the topic. These factors, while not necessarily indicative of educational poverty at an individual level, may, through their interactions, contribute to shaping local contexts in which an individual's right to full development is partially compromised.

A way to effectively disseminate the information entailed in such an extensive set of indicators is synthesizing the componing dimensions together [3, 9, 10, 13].

[1] The analysis had to abduct Aosta Valley and Trentino-Alto Adige because of a lack of data on five indicators (4 out of 7 in the drivers' dimension).

[2] There are two exeptions: non-profit sports institutions per 10,000 inhabitants 0–19 years old and cultural events per 100 inhab. 0–19 years.

To synthesize the individual indicators into a single measure, the Adjusted Mazziotta-Pareto Index (AMPI) is employed [8]. It serves as a partially non-compensatory composite indicator, standardized based on individual indicators at a reference time, thereby rendering the indicators independent of the unit of measurement. Consequently, all indicators are assigned equal weights, allowing for absolute time comparisons. Specifically, a re-scaling of the individual indicators within the range (70; 130) is proposed, defining two 'goalposts' that encapsulate the possible range of each variable across all time periods and units. The base value of 100 is derived from the reference municipality.

Two composite indexes have been calculated one for each dimension, outcomes and drivers. For both indices, Italy's score is equal to 100. Since indicators are build to identify possible scenarios of risk or deprivation scores, values above 100 report a situation relatively worse than the national average, while scores below 100 report a relatively better situation. We can not refer to them as poverty indicators because we did not work on setting a threshold. Hereinafter, we name the two composite as deprivation in outcomes and deprivation in drivers (or resources), and we use them to map the Italian territory, highlighting areas where intervention is more urgent.

4 Results

A joint analysis of the spatial distribution of the two indices reveals a moderate level of positive correlation between the two phenomena (correlation index 0.50). This result is not surprising considering the various factors that influence the relationship between drivers and outcomes, including those stemming from the educational community.

Interesting insights emerge from an analysis of the quadrants into which the scatter plot is divided. The upper-right quadrant groups the most disadvantaged areas, with above-average levels of deprivation in both resources and outcomes. We find here all areas of Sicily, Apulia, Campania, and many rural areas of some regions of the Center-North (Lazio, Liguria, Emilia Romagna) and of the South (Sardinia and Calabria). Finally, Friuli's cities and Basilicata's suburbs are also located in this quadrant, although their values are very close to the national average.

The regions in the opposite quadrant (lower left) show the best situation on both dimensions. We find here most of the cities of the northern and central regions, except those in Piedmont, Liguria, and Tuscany, which show slightly above-average values in terms of deprivation in outcomes, and the cities of Latium, which instead show above-average values concerning deprivation in drivers. We also find in this quadrant the cities of some southern regions, namely Abruzzo, Basilicata and Molise. The only rural area present is the one of Friuli Venezia Giulia. In the upper left quadrant (high deprivation in outcomes, low deprivation in drivers), in addition to the aforementioned cities of Piedmont, Liguria, and Tuscany, we find the anomalous positioning of the cities and urban suburbs of Sardinia. These cities show a very high level of deprivation in outcomes while facing low levels of deprivation in resources. The suburbs of Tuscany, Emilia Romagna, and Liguria are positioned in this quadrant, with values just above the national average regarding outcome deprivation.

In the lower right quadrant we find areas with below-average deprivation in outcomes despite low deprivation in drivers. Here, we find many rural areas, such as the cities of Latium, Calabria, and Apulia, as well as the suburbs of Lombardy (Fig. 2).

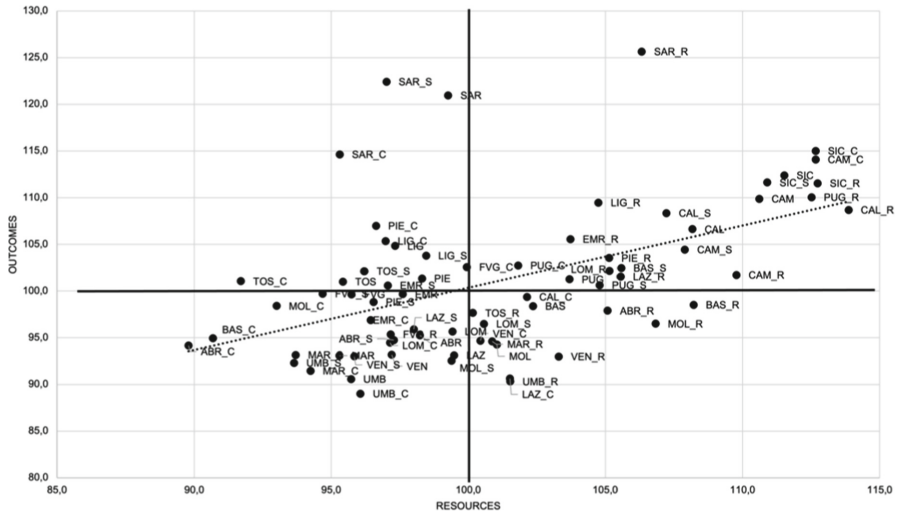


Fig. 2. Scatter-plot of the composite index of deprivation in drivers (x-axis) and the composite index of deprivation in outcomes (Source: Authors elaboration on multi-source data)

5 Conclusions

Despite being a preliminary exercise, some conclusions can be drawn that might help the Commission finalize its work.

Scouting for indicators to provide a quantitative measure to the theoretical framework has revealed important information gaps concerning certain themes (e.g., noncognitive skills), certain population segments (serious lack of information for the 0–5 year old age group), and the ability of the available sources to arrive at a sufficiently fine spatial detail (at least sub-provincial).

The degree of urbanization classification has proved useful in detailing a spatial pattern beyond the traditional north vs. south division. To be adequately interpreted, further analysis of the main determinants of the phenomenon is required.

To produce an educational poverty measure, a poverty line must be drawn. To calculate an individual educational poverty measure, data must be available for all indicators for the same person. If not, only a measure of poverty for a single geographical area can be calculated. The threshold must be set at indicator and composite index level; it can be deterministic or fuzzy, relative or absolute, normative or data-driven. The Commission needs to further elaborate on this topic.

References

1. Allmendinger, J., Leibfried, S.: Education and the welfare state: the four worlds of competence production. J. Eur. Soc. Policy **13**(1), 63–81 (2003)
2. Battilocchi, G.: Educational poverty in Italy: concepts, measures and policies. Cent. Eur. J. Educ. Res. **2**(1), 1–10 (2020)

3. Booysen, F.: An overview and evaluation of composite indices of development. Soc. Indicators Res. **59**, 115–151 (2002)
4. Botezat, A.: Educational poverty. NESET II ad hoc question n. 5 (2016)
5. Checchi, D.: Povertà ed istruzione: alcune riflessioni ed una proposta di indicatori. Polit. Econ. **14**(2), 245–282 (1998)
6. European Commission. Education and Training Monitor. Luxembourg: Publications Office of the European Union (2015)
7. Glaesser, J.: Relative educational poverty: conceptual and empirical issues. Qual. Quant. **56**, 2803–2820 (2022)
8. Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite indicator for measuring socio-economic phenomena. Soc. Indicators Res. **127**(3), 983–1000 (2016)
9. Mazziotta, M., Pareto, A.: Synthesis of indicators: the composite indicators approach. In: Maggino, F. (ed.) Complexity in Society: From Indicators Construction to their Synthesis. Social Indicators Research Series, pp. 159–191. Springer (2017)
10. OECD-JRC: Handbook on constructing composite indicators: Methodology and user guide. Paris: OECD (2008)
11. OECD: Measuring What Matters for Child Well-being and Policies. OECD, Paris (2021)
12. Pratesi, M., Quattrociocchi, L., Bertarelli, G., Gemignani, A., Giusti, C.: Spatial distribution of multidimensional educational poverty in italy using small area estimation. Soc. Indic. Res. **156**, 563–586 (2021)
13. Salzman, J.: Methodological choices encountered in the construction of composite indices of economic and social well-being. Center for the Study of Living Standards, Ottawa (2003)
14. Save the Children: La lampada di Aladino. L'indice di Save the Children per misurare le povertà educative e illuminare il futuro dei bambini in Italia. Save the Children Italia, Rome (2014)
15. Save the Children. Nuotare controcorrente. Povertà educative e Resilienza in Italia. Save the Children Italia, Rome (2018)

A Model for the Evaluation of the Italian University System

Ida Camminatiello(✉), Mario Pezzillo Iacono, and Rosaria Lombardo

University of Campania L. Vanvitelli, Capua, Italy
ida.camminatiello@unicampania.it

Abstract. Traditionally, the evaluation of university activities in Italy has been left to the discretion of individual degree programs, lacking a structured approach needed for such a critical task. However, significant changes in the Italian university system, have highlighted the need for systematic research programs that assess outcomes' quality effectively. Since 1999, the National Committee for the Evaluation of the University System has played a pivotal role in standardizing indicators and their definitions, enabling both vertical and horizontal comparisons. Additionally, several studies primarily focused on teaching activities and broader issues concerning university evaluation. In contrast, this research aims to analyze specific indicators influencing the quality of degree courses using an appropriate regression model.

Keywords: University evaluation · partial least squares regression · quality indicators

1 Introduction

The evaluation of university activity in Italy has traditionally relied on the discretion of individual degree courses (CdS), lacking the systematic approach necessary for such a critical objective. The significant changes in the Italian university system, transitioning from centralization to managerial and economic autonomy for individual universities, have heightened the demand for research programs that systematically assess the quality of outcomes. Formally introduced in 1989 (Law 168/89) with the establishment of the Ministry of University and Scientific and Technological Research (MURST), internal evaluation within universities gained prominence. However, specific evaluation units for university activities were established four years later (Law 537/93, Art. 5), outlining their primary responsibilities, including the annual report submission to the Ministry, the National University Council (CUN), and the Permanent Conference of Rectors of Italian Universities (CRUI). Centralized coordination was entrusted to the Observatory for the Evaluation of the University System, mandated to compile an annual comprehensive report for the Ministry based on data collected from individual university units. In 1999 (Law 370/99), the National Committee replaced the Observatory, coinciding with the renaming of the Ministry to

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 152–155, 2025.
https://doi.org/10.1007/978-3-031-64346-0_26

Education, University and Research (MIUR). Evaluation emerged as a strategic tool for monitoring goal attainment through systematic analysis of training, research, and management processes and outcomes. A critical initial step in any evaluation process is identifying indicators that define quality and guide assessment. An initial set of 80 indicators, spanning context, resources, process, and product areas, was derived from the analysis of data collected via a questionnaire administered to various university structures during the 1992/93 academic year. These indicators aimed to gauge the internal effectiveness of the university system in achieving predetermined objectives. Since 1999, the National Committee for the Evaluation of the University System has regulated and standardized the indicators and their operational definitions, facilitating vertical (within the same university) and horizontal (between different universities or CdS) comparisons. Additionally, the Center for Social Investment Studies (CENSIS) has annually produced rankings based on evaluations of faculties across different universities, though variations in indicators between academic years raise some concerns [4]. The National Committee for the Evaluation of the University System assesses both university-wide and degree-specific activities, enabling comparisons between these two levels. While earlier studies concentrated on teaching activities [2] and issues related to the evaluation of the university system [6], this research seeks to analyze indicators that influence the quality of CdS using an appropriate regression model. Considering that certain indicators exhibit high correlation, we opt for partial least squares (PLS) regression [5] as the preferred statistical methodology for handling this dependency in modeling [7] .

2 Evaluating the Quality of Degree Courses Trough PLS Regression

Selecting the dependent variable was a challenging process. Initially, we considered using the percentage of undergraduates who were overall satisfied with the CdS. However, the model developed did not yield a good fit. Therefore, we decided to measure dissatisfaction using the percentage of dropouts from the CdS after N+1 years labeled in Table 1. The predictors are detailed in Table 1 too. These indicators were gathered from 49 CdS at the University of Campania L.Vanvitelli, all of which are simple indicators.

The PLS regression run by the Durand's functions [3] in the R programming environment (see www.jf-durand-pls.com) carries out a model with two significant components that explain the 51.8 % of the regressors variability and the 45.3% of outcome variability with a $PRESS$ equal to 0.7012. To assess the relevance of the predictors we look at Table 2 that shows the corresponding PLS coefficients computed from two extracted components. These coefficients, derived from centered and scaled data, indicate the strength of the relationship between the dependent variable and all the independent variables in the model. The analysis reveals that the first and second predictor in order of importance are **iC13** (Percentage of CFU obtained in the first year out of the total CFU to

Table 1. The indicators considered in the regression model

Label	Indicator
iC01	Percentage of students enrolled within the normal duration of the CdS who have acquired at least 40 educational credits (CFU) during the academic year
iC05	Ratio of regular students to teachers
iC08	Percentage of permanent members be longing to basic and defining scientific-disciplinary sectors (SSD) for each CdS, of which they are the reference teachers
iC13	Percentage of CFU obtained in the first year out of the total CFU to be obtained
iC19TER	Hours of teaching delivered by permanent members and fixed-term researchers as a percentage of the total teaching hours
iC24	Percentage of dropouts from the CdS after N+1 years
iC25	Percentage of undergraduates who were overall satisfied with the CdS
iC27	Ratio of enrolled students to total teachers (weighted by teaching hours
iC28	Ratio of first-year enrolled students to first-year course teachers (weighted by teaching hours)

be obtained) and **iC01** (Percentage of students enrolled within the normal duration of the CdS who have acquired at least 40 CFU during the academic year), respectively. However, the estimates of their regression coefficients are negative indicating an inverse relationship with the response (i.e. the higher the indicators **iC13** and **iC01** are, the lower the **iC24** is). The third predictor is **iC28** (Ratio of first-year enrolled students to first-year course teachers) that has a positive effect on **iC24**. It follows **iC19TER** (hours of teaching delivered by permanent members and fixed-term researchers as a percentage of the total teaching hours) and **iC25** (Percentage of undergraduates who were overall satisfied with the CdS) which are negatively related to the response variable, meaning that as **iC19TER** and **iC25** increase, **iC24** (Percentage of dropouts from the CdS after N+1 years) tends to decrease.

3 Final Remark

In this paper, we explored the indicators that influence the percentage of dropouts from the CdS after N+1 years by modeling the dependency relationship using the PLS regression model. The selection of indicators for evaluating the university system has been a topic of great debate in the literature. A comprehensive discussion of the selection process and the effects of utilizing certain indicators over others in modeling [1] will be considered for future research. The

Table 2. Estimates of the PLS regression coefficients for the response variable.

Label	Estimates of the PLS regression coefficients
iC01	−0.249
iC05	0.037
iC08	0.062
iC13	−0.398
iC19TER	−0.124
iC25	−0.099
iC27	0.026
iC28	0.226

goodness of fit and the predictive ability of the proposed model are satisfactory but not excellent. Therefore, other shrinkage estimators or non-linear PLS methods could be considered. Finally, dissatisfaction has been measured using a simple indicator related to the percentage of dropouts from the CdS after N+1 years, and various determinants have been examined to explain and predict this indicator. It is particularly noteworthy that increasing the Percentage of CFU obtained in the first year out of the total CFU to be obtained and the Percentage of students enrolled within the normal duration of the CdS who have acquired at least 40 CFU during the academic year have been found to be highly relevant factors.

Acknowledgment. This research was sponsored by Italian Ministerial grants PRIN-2022 "SCIK-HEALTH" (code: 2022825Y5E_02; CUP: B53D23009750006) and PRIN-2022 PNRR "The value of scientific production for patient care in Academic Health Science Centres" (code: P2022RF38Y; CUP: B53D23026630001).

References

1. Alaimo, L.S., Ciacci, A., Ivaldi, E.: Measuring sustainable development by non-aggregative approach. Soc. Indic. Res. **157**, 101–122 (2021)
2. Bini, M., Chiandotto, B.: La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici. Studi Note Econ. **3**, 29–61 (2003)
3. Durand, J.F.: Local polynomial additive regression through PLS and splines: PLSS. Chemomet. Intell. Lab. Syst. **58**, 235–246 (2001)
4. Loreti, C.: Gli indicatori di valutazione del sistema universitario italiano: un'analisi critica. Quaderni Sociol. **48**(35), 81–102 (2005). https://doi.org/10.4000/qds.1112
5. Wold, H.: Partial least squares. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 6, pp. 581–591. Wiley, New York (1985)
6. Yorke, M: Performance indicators relating to student development: can they be trusted? Qual. High. Educ. **4**(1) (1998)
7. Camminatiello, I., Lombardo R., Musella, M., Borrata, G.: A model for evaluating inequalities in sustainability. Soc. Indicators Res. (2023). https://doi.org/10.1007/s11205-023-03152-3

Evaluation of the Quality of University Research from a Regional Perspective Using a Synthetic Indicator

Angela Maria D'Uggento(✉), Nunziata Ribecco, Vito Ricci, and Ernesto Toma

University of Bari Aldo Moro, Bari, Italy
angelamaria.duggento@uniba.it

Abstract. This paper deals with a proposal for the evaluation of scientific research in Italian universities from a regional perspective. A synthetic indicator based on the Wroclaw taxonomic method was constructed using the percentile rankings of 61 Italian universities in three notable international rankings and in the ANVUR-VQR assessment. Despite the simplicity of the methodological approach used to generate global rankings, we emphasize that a ranking is not able to capture the complexity of a multidimensional phenomenon that depends on both endogenous and exogenous variables of a university. Rankings are often criticized by academics but have a strong influence on stakeholders, potential students, lecturers, policy makers and funding bodies.

Keywords: Research university rankings · VQR · Wroclaw taxonomic method

1 Introduction

Universities are evaluated through many rankings [1], some of which specialize in research, while others cover overall performance, including teaching, services and third mission activities. These rankings serve multiple stakeholders, including potential students, faculty, policy makers and funding agencies. By providing a quantifiable measure of research excellence, rankings can influence decision-making, resource allocation and strategic planning. Prospective students and faculty often rely on research quality rankings as an indicator of an institution's academic caliber and reputation [2]. Highly ranked universities are perceived as centers of excellence and intellectual rigor and attract talented individuals who want to contribute to and benefit from such an environment [3]. These rankings enable students and researchers to make informed decisions about their educational and career paths. Policy makers and funding agencies can rely on rankings based on research performance when making their decisions about the allocation of public and private research funding [4]. Institutions that demonstrate high research quality are more likely to receive funding and attract potential collaborators, donors and partners in industry and academia. This undoubtedly creates a competitive environment that can foster excellence in research. For universities, research rankings provide a benchmark against which they can measure their performance against others.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 156–163, 2025.
https://doi.org/10.1007/978-3-031-64346-0_27

They can highlight strengths and areas for improvement, supporting strategic planning and investment in research infrastructure and talent [5].

However, research rankings are highly debated by academics. Some of them argue that relying on specific metrics can oversimplify the complexity of research quality and tempt universities to prioritize areas that improve rankings at the expense of broader academic missions [6]. Undoubtedly, the focus on quantifiable outcomes can disadvantage disciplines with different publication and citation practices. Vernon et al. [7] argue that rankings that rely heavily on subjective aspects such as reputation, award-winning faculty or alumni in high leadership positions do not effectively support the improvement of academic or research performance. We believe that improvement initiatives are needed and should focus on assessing the research performance of universities using comprehensive and standardized indicators. It is worth noting that rankings are considered reliable and trustworthy by their stakeholders and users, regardless of the methodology used to produce the indicators.

There are several notable rankings at the international level. Some of them provide an overall assessment of education and research activities, others are specifically focused on research: Best Global Universities Rankings (BGUR), SCImago Institutions Rankings with a focus on universities and University Ranking by Academic Performance (URAP). In Italy, the Italian Agency for the Evaluation of Universities and Research Institutes (ANVUR) is in charge of evaluating the research quality (VQR) of public and private universities and research institutes [8]. The Minister for Universities and Research establishes the VQR guidelines and provides the necessary funding for these assessments. The results of this timely review also determine the allocation of funds from the Ordinary Financing Fund (FFO), which supports the Italian university system [9]. This paper aims to contribute to the analysis of the quality of university research from a regional perspective.

2 Data and Methods

Among the international rankings based on the quality of research and the third mission, we selected those with the highest number of Italian public universities. As shown in Table 1, the following rankings proved to be comparable to the Italian VQR: BGUR, SCImago and URAP. Some Italian universities are also represented in the National Taiwan World University Ranking, however, this was not included in the analysis, as the small number of Italian universities would have led to a reduction in the number of rows in the overall matrix used, which currently contains 61 public universities. All rankings considered refer to the period 2015–2019, just like the ANVUR VQR data. Information on the indices used in each ranking can be found in Table 1. As to VQR data, the ministerial decrees provide a ranking of universities and departments based on sector-specific indicators, taking into account the size of the institutions in terms of expected products and researchers. Quality profiles for the products and two types of indicators, called R and IRAS, are used. The R indicator is qualitative and assesses the quality of the institution's products compared to the average, considering the importance of the different scientific fields. The IRAS indicator is qualitative-quantitative and evaluates the quality of the products, also considering the size of the institution.

Table 1. Indices and weights of selected international HEs research rankings that are comparable with the Italian VQR.

URAP		SCIMAGO		BGUR	
Indicators	W	Indicators	W	Indicators	W
Current Scientific Productivity	0,21	Normalized Impact (NI)	0,13	Global research reputation	0,125
Research Impact	0,21	Excellence with Leadership (EwL)	0,08	Regional research reputation	0,125
Scientific Productivity	0,1	Output (O)	0,08	Publications	0,1
Research Quality	0,18	Scientific Leadership (L)	0,05	Books	0,025
Research Quality	0,15	Not Own Journals (NotOJ)	0,03	Conferences	0,025
International Acceptance	0,15	Own Journals (OJ)	0,03	Normalized citation impact	0,1
		Excellence (Exc)	0,02	Total citations	0,075
		High Quality Publications (Q1)	0,02	Number of publications that are among the 10% most cited	0,125
		International Collaboration (IC)	0,02	Percentage of total publications that are among the 10% most cited	0,1
		Open Access (OA)	0,02	International collaboration – relative to country	0,05
		Scientific Talent Pool (STP)	0,02	International collaboration	0,05
		Innovative Knowledge (IK)	0,1	Number of highly cited papers that are among the top 1% most cited in their respective field	0,05
		Patents (PT)	0,1	Percentage of total publications that are among the top 1% most highly cited papers	0,05
		Technological Impact (TI)	0,1		
		Altmetrics (AM)	0,1		
		Inbound Links (BN)	0,05		
		Web Size (WS)	0,05		

A suitable combination of these indicators leads to a final IRFS indicator for each university. The overall IRFS indicator is defined as follows:

$$IRFSi = 0.90 * IRAS1\_2i + 0.05 * IRAS3i + 0.05 * IRAS4i \quad (1)$$

where IRAS1_2 is calculated by jointly considering permanent staff and new hires, IRAS3 is the indicator for research training and IRAS4 is the indicator related to the quality of research valorization.

To perform the analysis, a complex procedure was used that allowed us to collect the scores obtained by each of the 61 Italian universities in the 4 selected rankings and convert them into percentile ranks within the group of Italian universities (Table 2). The average percentile ranks were then calculated on a regional basis in the 4 rankings to obtain the vector of regional average percentile. These data were synthesized using the Wroclaw taxonomic method allowing us to distinguish groups of homogeneous regions [10, 11]. A map was created showing the quartiles of the distribution of Wroclaw indices by region. The data was retrieved from the websites in February 2024 and analyzed using the R software.

Table 2. Percentile ranks of Italian universities in the four rankings compared.

University	VQR	BGUR	SCI	URAP	University	VQR	BGUR	SCI	URAP
Ancona	40.0	36.7	41.7	46.7	Napoli Parthenope	30.8	28.3	13.3	16.7
Bari	83.3	68.3	71.7	78.3	Napoli Vanvitelli	66.7	51.7	46.7	61.7
Bari Politecnico	17.5	43.3	28.3	20.0	Padova	95.0	100.0	96.7	98.3
Basilicata	22.5	23.3	10.0	11.7	Palermo	85.0	45.0	63.3	68.3
Bergamo	27.5	0.0	0.0	1.7	Parma	61.7	58.3	61.7	55.0
Bologna	98.3	98.3	95.0	95.0	Pavia	70.0	83.3	81.7	76.7
Brescia	43.3	55.0	50.0	56.7	Perugia	71.7	73.3	78.3	71.7
Cagliari	64.2	35.0	40.0	48.3	Piemonte Orientale	33.3	15.0	45.0	30.0
Calabria	53.3	33.3	43.3	43.3	Pisa	88.3	90.0	88.3	88.3
Camerino	15.0	38.3	20.0	15.0	Pisa Normale	1.7	70.0	31.7	28.3
Cassino	10.0	6.7	6.7	5.0	Pisa S. Anna	5.0	31.7	51.7	18.3
Catania	76.7	65.0	75.0	73.3	Reggio Calabria	13.3	5.0	15.0	3.3
Catanzaro	11.7	13.3	30.0	26.7	Roma C.Biomedico	3.3	10.0	23.3	23.3
Chieti e Pescara	45.0	20.0	38.3	45.0	Roma La Sapienza	100.0	96.7	100.0	100.0
Ferrara	50.0	46.7	55.0	63.3	Roma Tor Vergata	81.7	80.0	85.0	80.0
Firenze	90.0	88.3	91.7	90.0	Roma Tre	60.0	48.3	26.7	36.7
Foggia	27.5	3.3	21.7	21.7	Salento	38.3	30.0	33.3	35.0
Genova	78.3	78.3	83.3	85.0	Salerno	68.3	63.3	58.3	66.7

(*continued*)

Table 2. (*continued*)

University	VQR	BGUR	SCI	URAP	University	VQR	BGUR	SCI	URAP
Insubria	30.8	18.3	25.0	38.3	Sannio	6.7	11.7	8.3	0.0
L'Aquila	35.0	8.3	35.0	40.0	Sassari	36.7	21.7	18.3	31.7
Messina	64.2	41.7	48.3	53.3	Siena	55.0	66.7	66.7	50.0
Milano	93.3	95.0	98.3	96.7	Torino	91.7	91.7	90.0	91.7
Mi Bicocca	75.0	85.0	80.0	86.7	Torino Politecnico	73.3	56.7	68.3	70.0
Mi Bocconi	25.0	53.3	1.7	10.0	Trento	51.7	86.7	70.0	65.0
Mi Cattolica	80.0	75.0	76.7	81.7	Trieste	48.3	71.7	56.7	51.7
Mi Humanitas	0.0	60.0	65.0	33.3	Tuscia	20.0	26.7	11.7	13.3
Mi Politecnico	86.7	81.7	86.7	83.3	Udine	46.7	40.0	36.7	41.7
Mi San Raffaele	8.3	76.7	73.3	75.0	Urbino	22.5	25.0	16.7	8.3
Modena R.E	57.5	50.0	53.3	60.0	Venezia	41.7	16.7	3.3	25.0
Molise	17.5	1.7	5.0	6.7	Verona	57.5	61.7	60.0	58.3
Na Federico II	96.7	93.3	93.3	93.3					

3 Results and discussion

All rankings show a strong positive and statistically significant correlation with the VQR (Table 3). This underlines the coherence of the measurements and the similarity of the methods used to assess research quality. This result was expected and confirmed our research hypothesis, as these rankings have refined their methodology over time to include a wider range of indicators capable of capturing the diversity of research excellence with a more complex methodology.

The map in Fig. 1 shows the regions labeled Q4, which are below the third quartile (i.e. the 25% with the lowest Wroclaw Index scores), the regions labeled Q3 with scores between the second and third quartile, the regions labeled Q2 with scores between the first and second quartile and the regions labeled Q1, which include the regions with the best universities on average from a research perspective.

The positive effects of the interaction between the quality of research and the economic situation of the region in which the university is located in terms of Gross Domestic Product, human resources, innovation and infrastructure are clearly visible. However,

Table 3. Spearman Rho correlation matrix between the percentile ranks of the 61 Italian universities

Ranking	VQR	BGUR	SCImago	URAP
VQR- Research Quality Assessment	1.000	0.741**	0.802**	0.890**
BGUR-Best Global Universities Rankings	0.741**	1.000	0.895**	0.876**
SCImago Insitution Rankings	0.802**	0.895**	1.000	0.952**
URAP-University Ranking by Academic Performance	0.890**	0.876**	0.952**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

some regions, such as Sicily, confirm that a regional higher education system can overcome an unfavorable economic situation with the efforts of its human capital and its global networks.

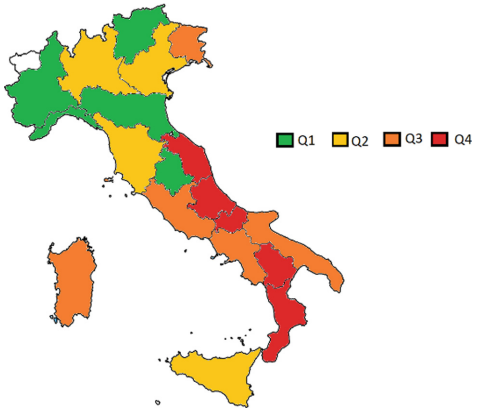


Fig. 1. Classification of the Italian regions according to the scores of the Wroclaw indices, which were calculated taking into account the universities located in each region.

4 Conclusions

Rankings of the research quality of universities are important tools that serve various stakeholders in the higher education ecosystem. They provide valuable insights into the research performance of institutions and promote competition and excellence in a common international framework. However, the responsible use of rankings requires an awareness of their limitations and a commitment to a holistic understanding of a multidimensional indicator such as research quality. To mitigate these problems, ranking organizations need to continuously refine their methodologies to include a broader range of

indicators that reflect the diversity of research excellence. Stakeholders should critically scrutinize the rankings and consider them alongside other qualitative assessments of the interactions of research with culture and territory, impact and innovation.

Although these aspects are inextricably linked, it is important to point out that the quality of academic education and research should be assessed with the awareness that it is influenced by structural differences between territories. While it is true that universities must be part of international networks, universities take from the local area contacts with the world of production, human resources, funding opportunities and, above all, students. All these relationships are part of the third mission, which, in addition to the two core tasks of research and teaching, is to promote the university's achievements in the region and make it a driving force for cultural and social development. The ANVUR evaluations point to opportunities for improvement for Italian universities as a whole and for the individual departments or scientific fields examined and highlight the existence of best practice in the Italian HE system. However, there are some significant differences in the final VQR ratings compared to international rankings. Understanding the reasons for these differences is the aim of future research, together with a proposal for the introduction of a standardized framework of common indicators in order to obtain more convergent results and a model that explores the relationships with macroeconomic regional indicators. An insight into the research findings of telematic universities is also in progress. The debate on the ability of rankings to provide comprehensive information, especially on research and educational outcomes, is far from over. Furthermore, the introduction of a more comprehensive set of indicators related to supervision, infrastructures and student evaluations should be considered to provide new information for a multidimensional concept such as the reputation of a university. Providing detailed information instead of rankings could be a better support for an informed decision-making process of stakeholders.

References

1. Finch, H.: An introduction to the analysis of ranked response data. Pract. Assess. Res. Eval. **27**, Article 7 (2022)
2. Horstschräer, J.: University rankings in action? The importance of rankings and an excellence competition for university choice of high-ability students. Econ. Educ. Rev. **31**, 1162–1176 (2012)
3. Hazelkorn, E.: Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence. Palgrave Macmillan (2015)
4. Marginson, S.: University rankings and social science. Eur. J. Educ. **49**(1), 45–59 (2014)
5. Aguillo, I.F., Bar-Ilan, J., Levene, M., Ortega. J.L.: Comparing university rankings. Scientometrics **85**(1), 243–256 (2010)
6. Cremonini, L., Westerheijden, D.F., Enders, J.: Disseminating the right information to the right audience: cultural determinants in the use (and misuse) of rankings. High. Educ. **59**(4), 465–483 (2010)
7. Vernon, M.M., Balas, E.A., Momani, S.: Are university rankings useful to improve research? A systematic review. PLoS ONE **13**(3), e0193762 (2018)
8. ANVUR. https://www.anvur.it/attivita/vqr/vqr-2015-2019/
9. Blasi, B., Romagnosi, S., Bonaccorsi, A.: Playing the ranking game: media coverage of the evaluation of the quality of research in Italy. High. Educ. **73**, 741–757 (2017)

10. Senetra, A., Szarek-Iwaniuk, P.: The implementation of the Wroclaw taxonomic method for the identification and evaluation of problem areas in the Warmia and Mazury Region in Poland – a case study. Socioecon. Plann. Sci. **67**, 43–57 (2019)
11. Delvecchio, F.: Scale di misura e indicatori sociali. Cacucci Editore, Bari (1995)

Causal STAR BART for Discrete Outcome

Dafne Zorzetto(✉)

Brown University, Providence, RI 02906, USA
dafne_zorzetto@brown.edu

Abstract. Accurate choices for modeling the probability distribution of variables are crucial in statistical analysis. However, in real-world applications, especially for discrete outcomes, these variables are often treated as continuous data, neglecting their inherent discrete nature. In this paper, we introduce a novel approach leveraging the Simultaneous Transforming and Rounding Process for Bayesian Additive Regression Trees (STAR BART) which is tailored to the causal inference framework. In contrast to well-known methods in causal inference, such as Causal BART and Bayesian Causal Forest (BCF), which are designed for continuous outcomes, an accurate adjusted STAR BART offers an alternative that explicitly addresses the challenges of discrete outcomes. Through a simulation study, we show the performance of our proposed approach in capturing heterogeneous treatment effects in datasets with discrete outcomes and compare it to Causal BART and BCF.

Keywords: causal inference · heterogeneous treatment effects · simultaneously transforming and rounding process

1 Introduction

In statistics, the modeling of probability distributions keeping into consideration the support of the variables is of paramount importance. However, real-world applications often encounter a significant challenge: the treatment of discrete outcomes as continuous data. A significant example arises in the potential outcome framework [12], where a flourishing literature is focused on defining models that are more and more flexible and able to capture the heterogeneity in the data. Significant examples are the increased use of the Bayesian Additive Regression Tree (BART) [2], extended for the causal inference framework in the Causal BART [6] and Bayesian Causal Forest (BCF) [5], and the dependent Dirichlet Process (DDP) mixture models [9,10] to overcame the fundamental problem of causal inference: the missing data problem [7] [see e.g., 1,11,13]. While these methods focus on capturing and characterizing the heterogeneity in the causal effect, they consider the outcomes in the continuous support.

In this work, we leverage the Simultaneous Transforming and Rounding Process (STAR) [8] to adjust the well-known Causal BART [6] to model discrete

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 164–169, 2025.
https://doi.org/10.1007/978-3-031-64346-0_28

potential outcomes. The proposed model (Causal STAR BART) allows us to impute the missing data taking into account their discrete nature and consequently estimate a coherent causal effect, without relinquishing to the flexibility of the model.

2 Causal Setup

Let i be the study unit, with $i \in \{1, \dots, n\}$, and $T_i \in \{0, 1\}$ be the binary treatment with observed value t_i. According to the Rubin Causal Model [12], the potential outcomes for unit i are defined as $\{Y_i(0), Y_i(1)\} \in \mathbb{N}^2$, for $i = 1, \dots, n$, where $\mathbb{N}$ indicate the set of positive discrete number that include 0. Invoking the Stable Unit Treatment Value Assumption (SUTVA) [12], $Y_i(0)$ is the outcome when the unit i is assigned to the control group, while $Y_i(1)$ is the outcome when it is assigned to the treatment group.

In practice, however, for $i = 1, \dots, n$, we observe only $y_i \in \mathbb{N}$, that is the realization of the random variable Y_i defined as

$$Y_i := (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1).$$

Conversely, we can not observe the realization $y_i^{mis} \in \mathbb{N}$ of the random variable Y_i^{mis} defined as $Y_i^{mis} := T_i \cdot Y_i(0) + (1 - T_i) \cdot Y_i(1)$.

Additionally, we define $\mathbf{x}_i \subseteq \mathcal{X}$ the p-dimensional vector of subject-specific background characteristics, covariates, and potential confounders—also called pre-treatment variables. Each vector $\mathbf{x}_i$ can contain both categorical and continuous variables. The tuple $(y_i, t_i, \mathbf{x}_i)$ for $i = 1, \dots, n$ therefore represents the observed quantities.

Our goal is to capture the heterogeneity in the causal effect of the treatment on the outcome, taking into consideration the discrete nature of the outcome. The causal estimand that can capture all the heterogeneity is the Individual Treatment Effect (ITE), defined as the difference between the two potential outcomes for each unit i

$$\tau_i = Y_i(1) - Y_i(0). \tag{1}$$

However, ITE cannot be observed and different causal estimands have to be considered. In literature considering observed covariates $\mathbf{x}$ to explain the heterogeneity in the data is well established [1,3,4]. Generally used is the Conditional Average Treatment Effect (CATE), defined as the expected value of the ITE, conditionally to observe the same particular values—or sets of values—for the covariates $\mathbf{X}$:

$$\tau(\mathbf{x}) := \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]. \tag{2}$$

To identify the causal effects, in addition to SUTVA the two following assumptions need to be invoked: (i) unconfoundedness—i.e., within subpopulations defined by values of observed covariates, the treatment assignment is random—and (ii) overlap–i.e., each unit i has the not null probability to receive both the treatment levels—such that

$$\{Y_i(1), Y_i(0)\} \perp T_i \mid X_i = x_i, \quad 0 < Pr(T_i = 1|X_i) < 1.$$

If the previous causal assumption holds, the statistical estimand of CATE (2) can be expressed as

$$\begin{aligned}\tau(\mathbf{x}) &= \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] \\ &= \mathbb{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}, T_i = 1] - \mathbb{E}[Y_i \mid \mathbf{X}_i = \mathbf{x}, T_i = 0]. \end{aligned} \tag{3}$$

3 Causal STAR BART

Due to the invoked causal assumptions and the statistical estimand for CATE (2), we can model the distribution of $\{Y_i \mid \mathbf{X}_i = \mathbf{x}, T_i = t\}$ for each $t \in \{0, 1\}$. However, the integer-valued nature of the outcome Y, suggests to use of a STAR process.

Following the definition of the STAR [8], we introduce the continuous variable W_i^*, for each unit $i \in \{1, \dots, n\}$, such that

$$Y_i = h(W_i^*),$$

where $h : \mathcal{H} \to \mathbb{N}$, with $\mathcal{H} \subseteq \mathbb{R}$, is the rounding operator, that guarantees the correct support for Y_i. The variable W_i operates as a continuous proxy for the count variable Y_i.

A second transformation through a strictly monotone function $q : \mathcal{H} \to \mathbb{R}$ introduces the latent variable Y_i^* for each unit i, such that

$$q(W_i^*) = Y_i^*.$$

The support of Y_i^* is more convenient for modeling following the common causal models.

Coherently with the causal framework, we define the probability distribution of the transformed outcome conditional to the treatment level and observed covariate as follows:

$$\{Y_i^* \mid \mathbf{x}_i, t\} = \mu_t(\mathbf{x}_i) + \epsilon_t(\mathbf{x}_i) \text{ for } t = \{0, 1\},$$

where $\mu_t(\cdot)$ is the conditional expectation of the transformed outcome and the errors $\epsilon_t(\cdot)$ are independent and normally distributed, such that $\epsilon_t(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2(\mathbf{x}))$. The error can be heteroscedastic, conditional on the covariates $\mathbf{X}$, and both the terms $\mu_t(\cdot)$ and $\epsilon_t(\cdot)$ are different by the treatment levels $t = \{0, 1\}$. Specifically,

$$\begin{aligned}\mu_0(\mathbf{x}_i) &= \mathbb{E}[Y_i^* \mid \mathbf{X}_i = \mathbf{x}, T_i = 0], \\ \mu_1(\mathbf{x}_i) &= \mathbb{E}[Y_i^* \mid \mathbf{X}_i = \mathbf{x}, T_i = 1].\end{aligned}$$

Among the model possibilities for $\mu_t(\cdot)$, we choose the BART [2] for the well-known feature of flexibility and ability to capture the heterogeneity in the data. Therefore, we can rewrite the conditional expectation of the transformed outcome as the sum of trees

$$\mu_t(\mathbf{x}) = \sum_{j=1}^{m} k(\mathbf{x}; A_j, M_j),$$

where A_j is the binary regression tree with terminal node parameters M_j.

4 Simulation Study

We investigate the performances of our proposed model through different simulated data settings and compare it with causal BART [6] and BCF [5].

For the following four scenarios, we simulate various continuous confounders $\mathbf{X}$, with standard normal distribution, and a binary treatment variable, such that $T_i \sim \text{Be}(\text{expit}(a(\mathbf{X})))$ for each unit $i \in \{1, \ldots, n\}$, where $a(\mathbf{X})$ is a linear function of the confounders $\mathbf{X}$.

Scenario 1: We simulate 15 confounders $\mathbf{X}$, that are also covariates for the outcome model, such that $Y(t) \sim Pois(b_t(\mathbf{X}))$, for $t = \{0, 1\}$. $Pois(\cdot)$ indicates a probability function of a Poisson distribution where the rate parameter is equal to $b_t(\mathbf{X})$. Due to the restriction for the support of the rate parameter, the function $b_t(\mathbf{X})$ is the absolute value of the linear regression of the covariates $\mathbf{X}$, where the parameters in it differ by the treatment level t.

Scenario 2: We investigate when in Scenario 1 we have 30 covariates that induce a big variability in the outcome distribution.

Scenario 3: We are interested in testing the model in the presence of a more complex structure of the outcome. With 15 confounders $\mathbf{X}$, the outcome distribution is a mixture of two Poisson probability functions, such that $Y(t) \sim \pi Pois(b_t(\mathbf{X})) + (1 - \pi)Pois(a_t(\mathbf{X}))$, for $t = \{0, 1\}$. The parameter π is equal to 0.7 and both the functions $b_t(\mathbf{X})$ and $a_t(\mathbf{X})$ the absolute value of the linear regression of the covariates $\mathbf{X}$.

Scenario 4: Correspond to Scenario 3 when the number of confounders increases to 30, and the variability increases as well.

The simulated scenario is reported in Fig. 1. The distributions of both the potential outcomes and the ITE (1) are more regular in the two scenarios than the distributions in scenarios 3 and 4. In particular, the mixture structure is well evident in scenario 3. Moreover, the variability increases in those scenarios where the number of confounders is double—i.e., scenarios 2 and 4.

We compare the performance of our proposed model—the causal STAR BART with identity transformation function—with the well-established models—causal BART [6] and BCF [5]. The results of the bias and mean square error (MSE) are reported in Fig. 2. As expected, the performance is similar for all the three models. However, our proposed model has slightly better results for the bias in all four scenarios, especially when the number of confounders increases. This is due to the STAR process that allows as to impute the missing data taking into consideration the discrete nature of the outcomes.

Fig. 1. Bar-plots of the four simulated scenarios. In the left, the distribution of the potential outcomes: the potential outcome under control in blue and the potential outcome under control in red. In the right the distribution of the ITE (1).

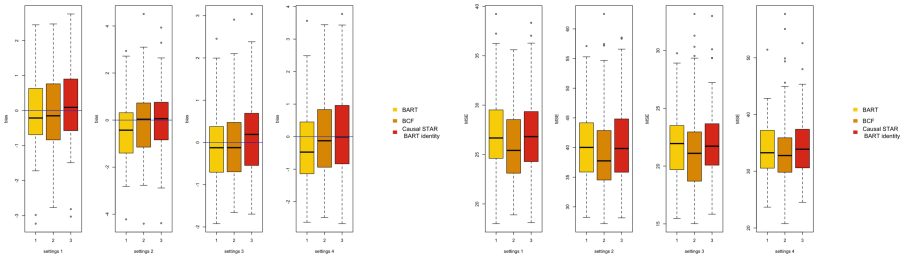


Fig. 2. Boxplot for the bias (left) and mean square error (MSE - right) for the average treatment effect for the four simulated scenario. results comparison of causal BART [6], BCD [5], and our proposed model (Causal STAR BART with identity transformation function).

References

1. Bargagli-Stoffi, F.J., Cadei, R., Lee, K., Dominici, F.: Causal rule ensemble: interpretable discovery and inference of heterogeneous causal effects. arXiv preprint arXiv:2009.09036 (2020)
2. Chipman, H.A., George, E.I., McCulloch, R.E.: Bayesian cart model search. J. Am. Stat. Assoc. **93**(443), 935–948 (1998)
3. Dahabreh, I.J., Hayward, R., Kent, D.M.: Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int. J. Epidemiol. **45**(6), 2184–2193 (2016)
4. Dwivedi, R., et al.: Stable discovery of interpretable subgroups via calibration in causal studies. Int. Stat. Rev. **88**, S135–S178 (2020)
5. Hahn, P.R., Murray, J.S., Carvalho, C.M.: Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal. **15**(3), 965–1056 (2020)
6. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)
7. Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986)
8. Kowal, D.R., Canale, A.: Simultaneous transformation and rounding (STAR) models for integer-valued data. Electron. J. Stat. **14**(1), 1744–1772 (2020)
9. MacEachern, S.N.: Dependent dirichlet processes. technical report. Department of Statistics, The Ohio State University, Columbus, OH (2000)
10. Quintana, F.A., Mueller, P., Jara, A., MacEachern, S.N.: The dependent Dirichlet process and related models. arXiv preprint arXiv:2007.06129 (2020)
11. Roy, J., Lum, K.J., Zeldow, B., Dworkin, J.D., Re III, V.L., Daniels, M.J.: Bayesian nonparametric generative models for causal inference with missing at random covariates. Biometrics **74**(4), 1193–1202 (2018)
12. Rubin, D.B.: Comment: which ifs have causal answers. J. Am. Stat. Assoc. **81**(396), 961–962 (1986)
13. Zorzetto, D., Bargagli-Stoffi, F.J., Canale, A., Dominici, F.: Confounder-dependent Bayesian mixture model: characterizing heterogeneity of causal effects in air pollution epidemiology. Biometrics (2024). In press

Data Science, Citizen Science and Smart Official Statistics

Elisabetta Carfagna[1(✉)], Gianrico Di Fonzo[2], Giovanna Jona Lasinio[3], and Paulo Canas Rodrigues[4]

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy
elisabetta.carfagna@unibo.it
[2] Sapienza University Roma and Italian Health Ministry, Rome, Italy
gianrico.difonzo@uniroma1.it
[3] Sapienza University Roma, Rome, Italy
giovanna.jonalasinio@uniroma1.it
[4] Department of Statistics, Federal University of Bahia, Salvador, BA, Brazil

Abstract. The discussion on advantages, disadvantages, limitations, and requirements of using alternative data sources integrated with probability sample surveys informs the debate in national and international statistical systems worldwide. The temptation to replace rigorous and costly data collection approaches with "smarter" ones is increasing. However, evaluating the reliability of statistics produced by elaborating alternative data sources is mandatory. In this work, we analyze the relationship between data science, new data sources, machine learning, citizen science and smart statistics, focusing on satellite data. We show that elaborating satellite data through parametric and machine learning classifiers does not always provide accurate statistics in complex landscapes, and machine learning classifiers do not systematically outperform parametric classifiers. Moreover, data collected by probabilistic samples play a crucial role. They should not be replaced by data collected by citizens without clear and strict guidelines in case statistics have to be produced.

Keywords: Data science · Citizen science · Sentinel satellites · Machine learning

1 Introduction

In today's digital era, the accessibility and analysis of extensive datasets from various sources, including administrative registers, satellites, social media, and digital platforms, have become increasingly feasible. [7] underscores the abundance of data in modern society, generated rapidly from diverse sources such as mobile phones, social media interactions, electronic transactions, and satellite imagery. These data sources present new opportunities for collection, processing, and storage, offering cost-effective alternatives to traditional statistical approaches involving surveys and questionnaires. However, this paper delves into critical questions regarding the reliability of statistics produced through data science applied to big data, exploring the relationship between data science, big data, new data sources, machine learning, and smart statistics. Additionally, the

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 170–175, 2025.
https://doi.org/10.1007/978-3-031-64346-0_29

paper investigates whether machine learning classifiers outperform parametric models and explores the potential of integrating machine learning with citizen-collected data to replace probability sample surveys.

2 The Data Set

A wide array of big data types, including administrative and unstructured data, are harnessed to generate statistics. However, unstructured data do not readily identify statistical units for integration with probability sample data, making its use for statistical purposes complex [8]. Relying solely on data science methods for processing such data can introduce significant biases that are difficult to quantify and mitigate. Satellite data, increasingly vital for analyzing agricultural and environmental phenomena, are based on raster-shaped pixels that may not align perfectly with territorial parcels, leading to complexities in classification tasks. This study uses a multitemporal satellite dataset covering a region north of the Tuscany Region, Italy, to evaluate its potential contribution to land cover estimation.

The dataset comprises six Sentinel 1 and Sentinel 2 images, six vegetation indexes for each of the Sentinel 2 images (Normalized Difference Vegetation Index (NDVI), Green Normalized Vegetation Index (GNDVI), Two-band Enhanced Vegetation Index (EVI2), Normalized Difference Water Index (NDWI), Chlorophyll Red-Edge (ClRededge), Soil-Adjusted Vegetation Index (SAVI)), and a Digital elevation model derived from satellite data. For the same area, the Italian Ministry of Agriculture kindly provided real ground data collected in 2016 on 574 geo-referenced points selected using a probabilistic sampling design in the framework of the AGRIT project. Information about land use cropping patterns, and farm management (soil cover, tillage practices, ground cover technique, irrigation, and presence of fences) was collected on the 574 points selected according to a two-phase probability area sampling strategy: a regular grid with 500 m side was overlaid on the territory in the first phase. The points at the cross of the grid were the first phase sample (aligned systematic sample in two dimensions) and were photo-interpreted on orthorectified aerial photos. Based on the photo interpretation, the points were attributed to the following land use strata: arable land, permanent crops, forage, scattered trees, forest and others. An additional stratification criterion was considered: low, medium, and high slope; thus, the intersection of the two stratification criteria generated the adopted stratification. The second phase sample (AGRIT sample) was a subset of the first phase, randomly selected according to specific sampling rates. The study area and the sample points are shown in Fig. 1.

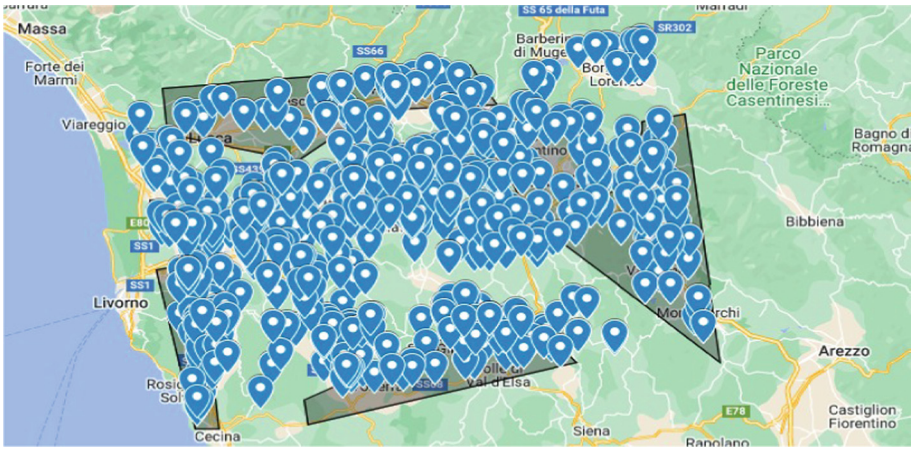


Fig. 1. Spatial distribution of the 574 s-phase sample points on which information about land use and farm management is collected. Dark areas contain a nonrandomly selected subset of the ground data (177 out) close to cities and coastal areas chosen to simulate citizen-provided data.

3 Performance of Parametric and Machine Learning Classifiers

[2] and [4] offer a comprehensive overview of the strengths and weaknesses associated with algorithms utilized for classifying satellite images. Linear classification methods are commonly employed for this purpose, with the parametric classification problem being addressed by partitioning the input space into regions labelled according to classification criteria [3]. In our study, we implemented the penalized logistic classifier, a classical parametric model, alongside three supervised machine learning classifiers [5]. The penalized logistic model employs likelihood maximization coupled with a lasso penalty term to manage a large number of explicative variables, selecting a subset of predictors while discarding the remainder. Additionally, we considered machine learning techniques such as bagging, random forest, and boosting. We implemented multiple splits in training and test sets, maintaining consistent proportions (1000 simulations). Specifically, 80% of the sample was allocated for training the classifiers (459 points), while 20% was reserved for testing (115 points). Boosting emerges as the most accurate classifier when ground data are incorporated among the explanatory variables, while random forest outperforms others when only remote sensing data are utilized. Penalized logistic multinomial regression ranks second in median accuracy, both with all explanatory variables and with satellite-derived variables alone. However, bagging exhibits the highest dispersion and lowest accuracy values among the classifiers assessed.

The achieved accuracies are notably lower when solely satellite-derived variables are considered. These findings contrast with prevailing trends observed in machine learning applications on remote sensing data, as reported in a 2019 review by [6]. These authors conducted a comprehensive review of machine/deep learning applications across various remote sensing datasets, particularly emphasizing publicly available benchmark datasets with extremely high resolution (pixel size < 10 m). However, such findings may not

fully reflect practical applications in complex landscapes, where satellite data classification serves as a proxy for crop acreage estimation, complementing ground data at the estimator level rather than offering a reliable estimate of crop acreage itself (Table 1).

Table 1. Median accuracy and Kappa value from the experiment. Classifiers are ordered according to their accuracy with all explanatory variables and satellite explanatory variables only

All explanatory variables			Satellite explanatory variables only		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.845	0.777	Random forest	0.691	0.526
Penalized logistic multinomial regression	0.836	0.767	Penalized logistic multinomial regression	0.689	0.521
Random forest	0.819	0.731	Boosting	0.677	0.528
Bagging	0.812	0.713	Bagging	0.652	0.495

4 Citizen Science Versus Probability Sample Surveys

In response to the growing trend of utilizing citizen-provided data and minimizing reliance on expert-led ground data collection efforts, we deliberately selected a non-random subset of 177 points from the original 574-point dataset. These points were strategically chosen to mimic the spatial distribution of data collected spontaneously by citizens, particularly focusing on areas near cities and coastal regions. From this subset, 142 points were allocated for training the classifiers, while the remaining 35 points were reserved for testing their accuracy. Additionally, to assess the impact of sampling methodology, we selected a stratified random subsample of 177 points from the original dataset. Due to the reduced sample size, the dispersion of accuracy across different training and test set splits is notably larger. This comparison enables an evaluation of classifier performance under varying sampling approaches, shedding light on the effectiveness of utilizing citizen-provided data instead of traditional ground data collection methods. The overall median accuracies and Kappa indexes of the classifiers with the probabilistic subsample and with the citizen sample are very similar (see Table 2). According to these results, a purposive sample of data collected by citizens seems to work as well as a stratified random sample for training and testing the classifier. This result can be misleading, since the purposive sample selection can generate a quite homogeneous sample and thus can facilitate the classification of the satellite images on sample units. However, when the trained classifier is applied to the whole study area, the result can be strongly biased.

In order to produce land use statistics, the classification of satellite data cannot be considered as the only data source, since pixel counting is known to be a biased estimator, particularly because of mixed pixels. Therefore, ground data are essential both for training and testing classifiers and for estimating the acreage of various land uses.

Table 2. Comparison of median accuracies and Kappa values for the various classifiers, using the subset of points close to cities and along the coast and the subset of points selected according to stratified random sampling.

Stratified random subsample			Citizen science subsample	
Classifier	Accuracy	Kappa	Accuracy	Kappa
Boosting	0.77	0.66	0.77	0.63
Penalized logistic multinomial regression	0.76	0.64	0.79	0.65
Random forest	0.72	0.59	0.74	0.55

Then, these estimates are improved combining ground-based estimates with classified data, treating the latter as auxiliary variables in a calibration or regression estimator [1].

To assess the possibility to replace the probability sample with citizen data, we compared the acreage estimates of different crops obtained from the non-probability sample with estimates derived from the entire AGRIT sample, consisting of the above mentioned 574 points selected via stratified random sampling. The expansion factor for the stratified random sample was determined based on the photo-interpreted systematic sample (29,658 points), and the same factor was applied to the estimates derived from the citizen subsample. The comparison presented in Table 3 reveals significant discrepancies in acreage estimates. Sunflower acreage is notably underestimated, while overestimations are observed for winter cereals, olive groves, and vineyards. It is important to note that in practical applications, expanding data to the entire population poses challenges, especially when collected by citizens without specialized skills and lacking a probability sample selection scheme. Consequently, the observed underestimations and overestimations may be even more pronounced in real-world scenarios.

Table 3. Difference in square kilometers and per cent between the acreage estimate of the various land uses based on the non-probabilistic sample design and the probabilistic sample design (AGRIT sample)

Vegetation Class	Area estimated with citizen subsample (ha)	Area estimated with AGRIT sample (ha)	Area estimated citizen-AGRIT	Relative difference
Other	148,654	152,364	−3,709	−2.43
Olive groves	52,896	47,043	5,853	12.44
Vineyard	50,577	37,368	13,209	35.35
Winter cereals	47,488	42,877	4,611	10.75
Sunflowers	5,285	9,07	−3,785	−41.73

5 Concluding Remarks

In this paper we have shown that the classification of Sentinel satellite data does not provide high accuracies in complex landscapes. We have also noticed that machine learning classifiers do not systematically outperform parametric classifiers. In fact, boosting is the most accurate classifier with the entire data set and all explanatory variables, random forest is the most accurate classifier with satellite explanatory variables only, and the penalized logistic multinomial regression has the second highest median accuracy both with all explanatory variables and with the explanatory variables derived from satellite data only. Furthermore, ground data collected on probability samples play a vital role in estimating land use acreage. In fact, the classification based on a purposive sample can be as accurate as the one based on a probability sample but can be affected by the bias of the non-probability sample. Thus, the land use estimates based on regression or calibration estimators with the classification as auxiliary variables are affected by the bias of the non-probability sample. It is important to note that expanding data to the entire population when collected by citizens who lack specialized skills and do not follow a probability sample selection scheme is problematic; consequently, observed discrepancies may be exacerbated in real-world applications.

References

1. Carfagna, E., Gallego, F.J.: Using remote sensing for agricultural statistics. Int. Stat. Rev. **73**, 389–404 (2005)
2. Defourny, P.: Land cover mapping and monitoring. In: Delincé, J. (ed.) Handbook on Remote Sensing for Agricultural Statistics (Chapter 2). Handbook of the Global Strategy to improve Agricultural and Rural Statistics (GSARS), Rome (2017)
3. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer Series in Statistics, vol. 1. New York (2001)
4. Gómez, C., White, J.C., Wulder, M.A.: Optical remotely sensed time series data for land cover classification: a review. ISPRS J. Photogramm. Remote Sens. **116**, 55–72 (2016)
5. Hamza, M., Larocque, D.: An empirical comparison of ensemble methods based on classification trees. J. Stat. Comput. Simul. **75**(8), 629–643 (2005)
6. Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A.: Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J. Photogram. Remote Sens. **152**, 166–177 (2019). ISSN 0924-2716
7. Pratesi, M.: Letter from the president. Surv. Stat. **88**, 4 (2023)
8. Tillé, Y., et al.: Some thoughts on official statistics and its future (with discussion). J. Off. Stat. **38**(2), 557–598 (2022). https://doi.org/10.2478/jos-2022-0026

Sampling Challenges from the 2030 Agricultural Sustainable Development Goals

Cristiano Ferraz(✉)

Department of Statistics, CASTLab - Computational Agricultural Statistics Laboratory, Federal University of Pernambuco, Recife, PE, Brazil
cferraz@castlab.org

Abstract. The 2030 Sustainable Development Agenda, proposed by the United Nations in 2015, involves efforts and strategies to achieve several global goals to fight poverty in all its forms. Several indicators were developed to keep track of the advances toward each of them. However, calculating these Sustainable Development Goal's (SDG) indicators may not be a trivial task, as they depend not only on the availability of data at each country, but also on complex measuring processes applied to sampling units selected by probability sampling designs. This paper presents a discussion on the main sampling issues found when dealing with estimating them, taking the 2.4.1 SDG indicator as an example. Conceptually defined as the proportion of agricultural land where productive and sustainable agriculture is practiced, this is a complex parameter to estimate. A general framework for estimating it under an agricultural dual frame design is presented, and the challenges involved in its estimation discussed.

Keywords: Dual-frame design · SDG indicators · Agricultural surveys

1 Introduction

The United Nation's (UN) 2030 Agenda on Sustainable Development encompasses 17 global goals, called the Sustainable Development Goals (SDGs), aiming at ending poverty in all its forms around the world. There are 169 targets related to the SDGs that need to be tracked by indicators, allowing to assess progress towards the goals. The United Nations Statistical Commission (UNSC), through the established Inter-Agency and Expert Group on SDGs (IAEG-SDGs), has developed a set of 232 Sustainable Development Goals' indicators. The UN's Food and Agriculture Organization (FAO) is the custodian of 21 of them [5], related to food, agriculture, and the sustainable use of natural resources. SDG indicators 2.3.1 and 2.4.1 are two of them.

Calculating SDG indicators is not a trivial task as they depend on countries' availability of data as well as on the measuring process involved in producing them. The two cited SDG indicators can be used as examples to illustrate the matter. The 2.3.1 indicator is defined as the volume of production per labor unit by classes of farming/pastoral/forestry enterprise size. The 2.4.1 indicator is defined as the proportion of agricultural land where productive and sustainable agriculture is practiced. The problem

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 176–180, 2025.
https://doi.org/10.1007/978-3-031-64346-0_30

of estimating each one of them is typical in the sense that the most cost-effective way to generate their estimates is using data collected through national agricultural statistics systems. In a general scenario, when the data source is an agricultural census, estimates of indicators rely on the availability of information for each agricultural farm. However, when data is collected through agricultural surveys, estimating SDG indicators depend also on the sample design.

There are many challenges related to the estimation process of SDG indicators based on survey sampling. For example, in the 2.3.1 SDG indicator case, there is the need for having data observed for special domains related to the classes of labor units. Requiring that over a national survey that already has several domains of interest based on geographical and administrative regions means almost surely to have to deal with estimation problems related to the small size of domains. In this context the capacity to explore different data sources based on satellite imagery may provide a way to cope with the problem [1]. While the 2.4.1 SDG indicator can also suffer from small domain's problems, it also has others estimation issues due to its own formulation [7]. For this reason, this paper focus on introducing the 2.4.1 SDG indicator and discussing its estimation under a typical agricultural survey based on a dual frame design. For a detailed discussion on estimating the SDG 2.4.1 indicator one can consult [2], and [6].

2 The 2.4.1 SDG Indicator

The 2.4.1 SDG indicator is directly related to Target 2.4 [5, 7], that looks to ensure sustainable food production systems and to implement resilient agricultural practices that increases both productivity and production. In a more concise way, the indicator is simply defined as the proportion of agricultural area under productive and sustainable agriculture, but its complexity goes beyond estimating a proportion. There are multi-dimensional characteristics involving economic, environmental, and social aspects reflected in its measurement process through a series of sub-indicators. Each sub-indicator corresponds to a proportion of productive and sustainable agricultural land according to a specific criterion. There are 11 of those criteria. SDG Indicator 2.4.1 is functionally defined as the minimum value of the sub-indicators, so to be interpreted as the largest portion of national agricultural land deemed to be sustainable according to all 11 criteria involved in its definition. Let SDG_{241} denote the 2.4.1 SDG indicator. In practice, it can be written as:

$$SDG_{241} = \acute{min}(R_{\#1}, \ldots, R_{\#11}), \tag{1}$$

where each $R_{\#c} : c = 1, \ldots, 11$, represents a sub-indicator defined by the expression:

$$R_{\#c} = \frac{TSAS\#c}{TSA}. \tag{2}$$

Equation (2) has $TSAS\#c$ representing a country's total agricultural area deemed to be productive and sustainable according to a given criterion c; The TSA represents the total national agricultural land area. Hence, a general estimator for the SDG_{241} can be defined as:

$$\widehat{SDG}_{241} = \acute{min}\left(\hat{R}_{\#1}, \ldots, \hat{R}_{\#11}\right), \tag{3}$$

where:

$$\widehat{R}_{\#c} = \frac{\widehat{TSAS\#c}}{\widehat{TSA}}. \tag{4}$$

Estimates based on (3) and (4) are complex to deal with due the functional forms involved. If on one hand, an estimator like (4) is a typical estimator for a population ratio, Eq. (3) shows that one must deal with 11 of these, at the same time it also must deal with an order statistic.

The measurement process related to the 2.4.1 indicator involves classifying each farm based on each criterion $c = 1, .., 11$, according to a set of three categories: green, yellow, and red [2]. Farms classified into a green category related to a criterion c are considered to have a desirable level of sustainability according to that criterion. Farms classified in the yellow category are considered to have an acceptable level of sustainability, while those classified into the red category are deemed unsustainable. Using this unified classification structure one can rewrite the SDG_{241} estimator (1) as a function of sub-indicators for each classification category. Considering an example based on the green category, it is possible to write:

$$R_{\#c} = \frac{\sum_{k \in U} y_k V_{(c)k}}{\sum_{k \in U} y_k}, \tag{5}$$

so that: $U = \{1, \ldots, N\}$ represents the universe of N agricultural farms of the country; $V_{(c)k}$ is an indicator function assuming value 1 when farm k is classified into the green category for the c criterion ($c = 1, \ldots, 11$), and zero otherwise; and y_k is the agricultural area reported by farm k. Therefore, a consistent estimator $\widehat{SDG}_{241}$ can be written as a function of the following sub-indicators when related to the category of desirable sustainability:

$$\widehat{R}_{\#c} = \frac{\sum_{k \in S} \frac{y_k V_{(c)k}}{\pi_k}}{\sum_{k \in S} \frac{y_k}{\pi_k}}, \tag{6}$$

where S represents the set of farms selected for the sample using a probability sample design so that π_k is the first-order inclusion probability.

3 Sampling Design and Estimation

Specific forms of each sub-indicator estimator $\widehat{R}_{\#i}$, and therefore the $\widehat{SDG}_{241}$ estimator itself, depend on the sample design employed in each country. In a general framework it is likely that several sampling elements are used in an agricultural survey, such as stratification, clustering, and unequal inclusion probabilities. In addition, a dual frame design can be adopted, using an area frame and a list frame. Without loss of generality, consider the problem of producing estimates for a stratum, and so notation regarding stratification is omitted. Also, suppose the element of analysis, the farm, coincide with the last level sampling unit. This scenario can accommodate typical agricultural surveys using a dual frame design based on an area frame and a list frame [3]. On one hand,

considering the design related to the list frame, three situations can happen: (i.) the list frame identifies all N farms of the population and a sample of n of them ($n < N$) is taken using probability sampling; (ii.) the list frame identifies the largest farms, and a census is carried out selecting all of them; and (iii.) the list frame identifies the largest farms, and a sample of them is taken using probability sampling. On the other hand, when considering the area frame design, there are many possibilities. Considering the use of an area frame of squared segments, estimation could be done using, for example, (a) an open segment strategy, (b) the one-stage weighted segment strategy, and (c), using a two-stage weighted segment strategy where farms are selected by points [8]. Decision to the proper design for the area and the list frame involves assessing the reality of each country. To illustrate, consider the general case where the area frame of square segments is used to select farms by points, in a two-stage sampling design [4], and the list frame is used to select a sample of large farms. In this scenario each sub-indicator estimator could be built based on the simple multiplicity dual frame estimator [9] given by:

$$\widehat{R}_{\#c} = \frac{\sum_{i\in S_A}\sum_{j\in U_i}\sum_{k=1}^{P}\frac{y_{ij}V_{(c)ij}I_{ijk}}{\pi_i\pi_{j|i}m_{ij}^A} + \sum_{k\in S_L}\frac{y_kV_{(c)k}}{\pi_k m_k^L}}{\sum_{i\in S_A}\sum_{j\in U_i}\sum_{k=1}^{P}\frac{y_{ij}I_{ijk}}{\pi_i\pi_{j|i}m_{ij}^A} + \sum_{k\in S_L}\frac{y_k}{\pi_k m_k^L}}, \tag{7}$$

where, from the area frame components: S_A represents the set of squared segments selected from the area frame; U_i represents the set of all farms that can be sampled from square segment i, k represents one of the P points used within a squared segment so select farm; π_i represents the inclusion probability of square segment i; $\pi_{j|i}$ represents the probability of selecting farm j given the square segment i was selected; I_{ijk} is an indicator variable that farm j was observed at point k, within square i; y_{ij} is the area reported by farm j selected through square segment i; $V_{(c)ij}$ is the indicator variable that farm ij was classified in the green category; and m_{ij}^A is the multiplicity factor associated to farm j selected through square segment i. Concerning the list frame components: S_L represents the set of farms selected from the list frame; π_k is the inclusion probability of a farm; y_k is the area reported by farm k, $V_{(c)k}$ is the indicator variable that farm k was classified in the green category; and the m_k^L is the multiplicity factor associated to farm k. Such a richness of notation reflects the many sampling design aspects needed to be properly considered when facing SDG indicator's estimation.

4 Concluding Remarks

The fact that the 2.4.1 SDG indicator estimator, defined by (3) and (4), needs to be further derived, to reflect the actual agricultural sampling design in use, such as exemplified in Eq. (7), is one of the issues that need to be addressed, when facing estimation of SDG indicators, but it is not the only one. The measuring process itself, where each farm needs to be classified in each category of green, yellow, or red is not simple [7]. Furnishing disaggregated estimates also faces challenges such as the need for estimation of small domains. However, proposing a small area estimation model for an SDG indicator such as the 2.4.1 is not trivial. The last issue mentioned in this short paper relates to the problem of assessing the statistical quality of the 2.4.1 SDG estimate. Variance estimation of such estimator is still a subject that needs development.

References

1. Clara, A.K., Di-Candia, S., Falorsi, P.D., Gennari, P.: Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators: a practical application on SDG indicator 2.3.1. Stat. J. IAOS **38**, 879–891 (2022)
2. Ferraz, C. Sampling and estimation guide for SDG Indicator 2.4.1 under multiframe designs. FAO Statistics Working Paper Series, No. 24–40. Second Edition. Rome, FAO (2024)
3. Ferraz, C., Mecatti, F., Torres, J.: Dual frame design in agricultural surveys: reviewing roots and methodological perspectives. Stat. Methods Appl. **32**, 593–617 (2023)
4. Gallego, F.J., Delincé, J., Carfagna, E.: Two-stage area frame sampling on square segments for farm surveys. Surv. Methodol. **20**(2), 107–115 (1994)
5. FAO. Tracking progress on food and agriculture-related SDG indicators 2023. Rome (2023)
6. FAO. Sampling guidance for SDG Indicator 2.4.1. Rome (2023)
7. FAO. Proportion of agricultural area under productive and sustainable agriculture. Methodological note, Revision 11. Rome (2023)
8. FAO. Handbook on master sampling frame for agricultural statistics: frame development, sample design and estimation. Rome (2015)
9. Mecatti, F.: A single frame multiplicity estimator for multiple frame surveys. Surv. Methodol.Methodol. **33**(2), 151–158 (2007)

Methodological Perspectives in Integration of Data from Multiple Probabilistic and Non-probabilistic Sources

Pier Luigi Conti(✉)

Sapienza Università di Roma, Rome, Italy
pierluigi.conti@uniroma1.it

Abstract. Among the main consequences of the rapid growth of available data proper of the data deluge era, a major one is a significant diversification of primary data sources. This has determined a change of the traditional paradigm of a unique main source of data coming from an *ad hoc* statistical data collection process, to a multi-source world where data from sources with different characteristics have to combined together. New problems coming from the multi-sources problem, and corresponding methodological challenges, are analyzed.

Keywords: Data integration · statistical matching · uncertainty · nonprobability samples

1 Introduction

In the last two decades, a rapid growth of different types of data sources (Official Statistics survey data, administrative data, experimental data, observational data, data from sensor, transactions data, etc.) has been experienced. Such data are used in different contexts: science, economics, commerce, Official Statistics, etc. An important aspect is that they are collected, at different levels of granularity, by different organizations and for different purposes. In several cases, they are in the public domain, or anyway available upon request. This is a part of the *Big Data Revolution*, that has opened new opportunities for researchers to access data potentially useful to investigate relationships among variables.

In the new Big Data era, *data integration* represents a major challenge. In general, the concept of data integration takes on different shades depending on the adopted methodological perspective. The traditional paradigm of a unique main source of data coming from an ad hoc statistical data collection process is rapidly changing. As a matter of fact, there is an increasing availability of data not only from traditional sources (such as Official Statistics survey data, administrative data, data from experimental studies, data collected in observational large-scale studies), but also from tracking online activities or real-time monitoring (transaction data, sensor data, etc.). Data of potential interest are

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 181–186, 2025.
https://doi.org/10.1007/978-3-031-64346-0_31

not only collected by National Statistical Agencies through sample surveys (for instance, the National Health Interview Survey - NHIS in USA), or contained in administrative registers kept by Administrative Bodies, but also come by interactions of individuals with the World Wide Web by using apps, Web browsers or social media sites, or by sellers (commercial transaction records, etc.); cfr. [1]. Handling big data in social sciences may be challenging, as well, to improve information retrieved from large surveys [5]. At any rate, end-users are experiencing a rapid growth of different types of data sources, freely or cheaply available and of potential interest and impact in many fields.

The concept of big data - commonly defined by volume, variety, velocity, and veracity - is considered here as strictly connected to the importance of a rigorous approach to data integration. This perspective is far beyond the simple construction of data platforms and data analytics, but includes statistical features of sampling, design, and hypothesis testing. Only within this perspective we can face the dual role of data integration, that is innovating the accessibility to huge amount of data as well as rooting out the error of thinking that big data automatically translates into big information and reliable inference.

2 Combining Data from Different Sources

As outlined in the Introduction, the traditional, well-established paradigm of statistics is that high quality data come from a (frequently *ad hoc*) statistical sample survey or experimental study, and are collected in an appropriate database. Statisticians use data to make inference on population parameters, and the sources of statistical error are two.

- Sampling error, due to the discrepancy between sample and whole population.
- Non-sampling errors, due to non-responses and measurement errors.

Within this paradigm, the main effort of statisticians is to reduce the magnitude of the statistical error by carefully studying two basic ingredients.

(i) Design of a good data collection process, either an experimental study or a sampling plan.
(ii) Development of good statistical methods to analyze collected data, that require to account for both sampling and non-sampling errors.

Even if the effect of non-responses and, more generally, of missing data could be dramatically relevant, and could potentially produce high biases, the traditional paradigm of statistical inference allows us to use, in a standard way, well established tools such as Mean Squared Error of estimators, coverage level of confidence intervals, p-value and power function of statistical tests. The use of such basic tools lies on an underlying assumption: each observation refers to an actually existing population unit. Data could be incomplete, or even empty, but they necessarily refer to real population units.

Although this traditional paradigm is of fundamental importance even now (it can be considered as an ever green), it suffers of non-negligible limitations, shortly listed below.

- *Cost*: *ad hoc* statistical studies are expensive, so that is it natural to resort, when possible, to alternative, already existing data sources.
- *Time*: the data collection process can be time-consuming, and it would be appealing taking data from already existing databases.
- *Nonresponses* and *refusal to participate* are a common problem, with steadily decreasing response rates to official surveys.

As a consequence, combining data from different sources, *i.e. data integration*, is becoming more and more important for several purposes, that can be roughly considered as falling into two categories.

1. Construction of new databases by combining together single ones, in order to have a broader domain of application.
2. Making statistical inference from several databases simultaneously considered.

The paradigm based on data integration is very promising, because of its potential advantages for scientific investigations. On the practical side, combining data from different sources is becoming more and more important either to construct new databases broadening the domain of application of single databases combined together, or to make statistical inference on several variates that are not jointly observed in a single statistical database. Therefore, there is now an increasing interest for data to be collected *via* integration.

On the theoretical side, the change of the traditional paradigm of a unique main source of data coming from an *ad hoc* statistical data collection process has relevant effects. A common feature of data coming from different sources is that common variables could be observed on sample units selected with different, possibly unknown, sampling designs, and by using different measurement methodologies, possibly with different levels of precision. To give an example, according to [11], consider a population (either finite or infinite), on which three multidimensional variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are defined. Let A, B, C be samples corresponding to different data sources. The main patterns occurring in combining data from different sources can be summarized in Table 1 (cfr. [11]).

The simplest missing pattern is the monotone one. It is typical, for instance, of a small sample A collected through a well-designed sampling plan where $\mathbf{X}$, $\mathbf{Y}$ are observed, and a large sample B, obtained by self-selection of units, where also variable $\mathbf{Z}$ is observed. In this case, the integrated use of the two samples is based on the idea of using sample A to adjust sample B in order to reduce bias due to self-selection.

In non-monotone missing pattern I different samples are considered, but one of them contains observations of all variables of interest.

Finally, non-monotone missing pattern II is the most difficult one, because there is no sample with complete observations. Among the main problems of data integration falling into this category, two are of particular interest: Record Linkage and Statistical Matching.

In Record Linkage [7], aka Entity Resolution or De-Duplication, two samples A, B, are available. They are composed by n_A, n_B (generally independent)

observations of $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Z})$, respectively, corresponding to records of datafiles containing sample data. The goal is to identify and link together the records of different datafiles corresponding to the *same statistical units* (*entities*). In each datafile, records are usually identified via *key variables*, containing common identifying information. However, a unique, error-free identifier is missing because not available. In the notation of Table 1, key variables are some of the $\mathbf{X}$-variates. Hence, linkage is usually characterized by uncertainty, mainly for the presence of possible *noise* in key variables.

Statistical Matching, that also fall into the non-monotone pattern II, but which is, in a sense, "opposite" to Record Linkage, will be dealt with in Sect. 3.

Intuitively speaking, combining together sample data can be interpreted as a problem of "filling the gap" due to missingness, namely to the lack of complete observations. Within the "new" paradigm of data obtained by combining different databases, there are new potential sources of errors. They are essentially related to a question: *"May we consider combined data as observations from the population they are supposed to represent?"* In [12], this is referred to as *entity ambiguity*

Statistical data obtained by combining partial observations from different sources do not necessarily correspond to observations of real units, because of the intrinsic *uncertainty* in the combination process. Combined partial observations actually correspond to *virtual units* that define, in their turn, a *virtual population*. As a result, data obtained by combination of different sources can be seen, in the best case, as a sample of the virtual population. Since the main object of interest are parameters of the real population, the possible discrepancy between the virtual population and the real one is a major drawback in statistical inference process, because it corresponds to an *additional source of uncertainty*.

In response to the new paradigm described above, the development of a framework for the analysis of data obtained through combination of partial observations from different sources is necessary. As above remarked, in addition to the sampling and non-sampling "traditional" errors, we have to consider specific errors for combined data, depending on the combination process. Apart a few very special cases, ignoring the possible discrepancy between virtual and real population, *i.e.* dealing with integrated data as if they were single-source data, and using standard methods for statistical analysis, is generally incorrect. The development of inferential methods accounting for all sources of uncertainty is therefore necessary. Generally speaking, the problems to be considered with integrated data obtained by combining different sources are two.

- Development of methodologies for data integration allowing for a safe assessment of different sources of errors. This problem arises when one has to decide *how to combine* data from different sources.
- Use of *secondary data*, *i.e.* data already obtained by some integration process. This is the case of secondary analysis, cfr. [10]. In this case, focus is in modeling errors and analyzing their impact on statistical inference.

3 Statistical Matching

As already said, Statistical Matching falls into the non-monotone pattern II. Consider a population (either finite or infinite), on which three variables $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are defined. In the sequel, we will assume that $\mathbf{X}$ is P-dimensional, $\mathbf{Y}$ is Q-dimensional, and $\mathbf{Z}$ is R-dimensional, respectively. Let A, B be two independent samples of size n_A, n_B, respectively. In sample A, only variates $\mathbf{X}$ and $\mathbf{Y}$ are observed and in sample B only variates $\mathbf{X}$ and $\mathbf{Z}$ are observed. $\mathbf{X}$ is the set of variates *common* to the two samples, whilst $\mathbf{Y}$, $\mathbf{Z}$ are the variates *specific* of sample A, B, respectively. No joint observation of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ is available.

The goal of *statistical matching* is twofold.

- At a *micro* level, statistical matching aims at constructing, for the $n_A + n_B$ sample units, a unique, synthetic database containing $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ values. Missing values in the two samples are *imputed*.
- At a *macro* level, statistical matching aims at estimating the joint distribution of ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$), or parameters related to such a distribution. The goal is to learn something on the relationship of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ variates (or only of $\mathbf{Y}$ and $\mathbf{Z}$), even if they are never jointly observed.

The macro and micro approaches to statistical matching are formally different but equivalent; cfr. [6].

Denote by $\hat{z}^A_{a1}, \ldots, \hat{z}^A_{aR}$ $(a = 1, \ldots, n_A)$ the imputed z-values in sample A, and by $\hat{y}^B_{b1}, \ldots, \hat{y}^B_{bQ}$ $(b = 1, \ldots, n_B)$ the imputed y-values in sample B, respectively. Each row of the synthetic database corresponds to a virtual (sample) unit, because either z-values or y-values are not really observed, but imputed. As a consequence, neither the joint distribution of $(Y^A_{a1}, \ldots, Y^A_{1Q}, X^A_{a1}, dots, X^A_{1P}, \hat{Z}^A_{a1}, \ldots, \hat{Z}^A_{aR})$ nor that of $(\hat{Y}^B_{b1}, \ldots, \hat{Y}^B_{bQ}, X^B_{b1}, \ldots, X^B_{bP}, Z^B_{b1}, \ldots, Z^B_{bR})$ (*i.e.* the virtual population distribution) do coincide with that of ($\mathbf{Y}$, $\mathbf{X}$, $\mathbf{Z}$). The discrepancy between those two distributions, in this setting, is the *matching noise*. A study of the matching noise for some statistical matching techniques, and of its asymptotic behaviour under special conditions is in [8]. The use of multiple imputation, and a proposal for appropriate sampling weights to be used for the virtual units of the combined samples, is in [9]. The main problem is that it requires the knowledge the actual sample weights of units in both samples A, B, and this information is hardly ever available in practice. Other references to this problem are in [2].

As far as the macro approach to statistical matching is concerned, unless special assumptions are made, the statistical model for the joint distribution of ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$) is *not identifiable*; cfr. [2,3]. Cfr. also [4], where the problem of multivariate matching is approached through Bayesian Networks.

Table 1. Main patterns occurring in data integration

Sample	X	Y	Z
Monotone missing pattern			
A	observed	observed	unobserved
B	observed	observed	observed
Non-monotone missing pattern I			
A	observed	observed	unobserved
B	observed	unobserved	observed
C	observed	observed	observed
Non-monotone missing pattern II			
A	observed	observed	unobserved
B	observed	unobserved	observed

References

1. Citro, C.F.: From multiple modes for surveys to multiple data sources for estimates. Surv. Pract. **40**, 137–161 (2014)
2. Conti, P.L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. J. Am. Stat. Assoc. **111**, 1715–1725 (2016)
3. Conti, P.L., Marella, D., Scanu, M.: How far from identifiability? A systematic overview of the statistical matching problem in a nonparametric framework. Commun. Stat. - Theory Methods **46**, 967–994 (2017)
4. Conti, P.L., Marella, D., Vicard, P., Vitale, M.: Multivariate statistical matching using graphical modeling. Int. J. Approximate Reasoning **130**, 150–169 (2021)
5. Dalla Valle, L., Kenett, R.: Social media big data integration: a new approach based on calibration. Expert Syst. Appl. **111**, 76–90 (2018)
6. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching - Theory and Practice. Wiley, Chichester (2006)
7. Herzog, T.N., Scheuren, F.J., Winkler, W.E.: Data Quality and Record Linkage Techniques. Springer, New York (2007)
8. Marella, D., Conti, P.L., Scanu, M.: On the matching noise of some nonparametric imputation procedures. Stat. Probab. Lett. **78**, 1593–1600 (2008)
9. Rubin, D.: Statistical matching using file concatenation with adjusted weights and multiple imputations. J. Bus. Econ. Stat. **4**, 87–94 (1986)
10. Vartanian, T.P.: Secondary data analysis. New York (NY), Oxford (2011)
11. Yang, S., Kim, J.K.: Statistical data integration in survey sampling: a review. Jpn. J. Stat. Data Sci. **3**(2), 625–650 (2020). https://doi.org/10.1007/s42081-020-00093-w
12. Zhang, L.-C., Chambers, R.L. (eds.): Analysis of Integrated Data. CRC Press, Boca Raton (2019)

An Information Theory Approach to Assess Residential Segregation: The Case of Messina (Italy)

Francesca Bitonti[1(✉)], Daniela Ghio[2], Angelo Mazza[3], and Massimo Mucciardi[4]

[1] University of Catania, 95129 Catania, Italy
francesca.bitonti@unict.it
[2] University of Catania, 95131 Catania, Italy
daniela.ghio@unict.it
[3] University of Catania, 95128 Catania, Italy
a.mazza@unict.it
[4] University of Messina, 98121 Messina, Italy
massimo.mucciardi@unime.it

Abstract. Residential segregation among immigrant populations in urban areas is a multifaceted phenomenon with significant social implications. This paper investigates the residential segregation of four primary immigrant groups—Sri Lankans, Filipinos, Romanians, and Moroccans—within the metropolitan area of Messina, Italy, in 2016. Leveraging anonymized individuals' data from the population register, the study employs Shannon's entropy and Kullback-Leibler (KL) divergence to quantify segregation and divergence from the native Italian population. Preliminary results suggest that while certain immigrant groups exhibit spatial concentration in terms of the Shannon's index, their distributions may not necessarily diverge substantially from the local population, as for the KL divergence. The disparities detected by KL divergence suggest that immigrants may interact and share environments with the local population, challenging simplistic assumptions about segregation. This study contributes to public debate giving insight into the complexity of residential segregation dynamics to promote social cohesion policy accounting for the specificities of settlement patterns in urban environments.

Keywords: Residential Segregation · Entropy · Spatial Analysis · KL divergence

1 Introduction

Residential segregation, as a multidimensional phenomenon [1], has long been analyzed through the lens of the information theory with concepts of information and its counterpart, entropy. Entropy, a concept frequently employed in physics and information theory, serves to gauge the level of randomness within a system or the informational content of a message [2–5]. Initially introduced by Theil [5, 6] into the realm of social sciences, entropy emerged as a metric for quantifying population diversity and income

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 187–192, 2025.
https://doi.org/10.1007/978-3-031-64346-0_32

inequality [7, 8]. In essence, entropy (here interpreted as a measure of uncertainty) represents the quantity of information required to delineate a probability distribution. Social researchers use entropy as a measure of residential segregation because it provides a way to quantify the degree of homogeneity within neighbourhoods or regions in terms of certain characteristics such as ethnicity, income, or education level. In the context of residential segregation, entropy becomes high when the probability of occurrence of two groups of individuals is equally likely over space, and decreases when one group has a higher probability compared to the other because there is less uncertainty about the outcome will occur [3].

Utilizing data sourced from the population register, this paper investigates the residential segregation patterns of migrants within the metropolitan area of Messina, Italy. The study offers a thorough analysis of segregation among the four primary immigrant groups—Sri Lankans, Filipinos, Romanians, and Moroccans—in 2016. The primary inquiries center around comparisons in the distribution of these groups, measuring (1) ethnic residential segregation exploiting the concept of entropy, (2) the divergence between immigrant group and natives' distribution. The paper is organized as follows: the second section presents the area under study together with the analyzed data and the methodology implemented. The third section illustrate the results of the application and the last one gives the final considerations and conclusions.

2 Study Area, Data and Methods

The geographical area of interest corresponds to the metropolitan area of Messina, located on the Northeastern coast of Sicily (Italy). Messina was almost destroyed by an earthquake in 1908, after which the city underwent a serious economic crisis [9]. After the seismic event, the city developed as a polycentric urban area which, through time, has fragmented the socioeconomic identity of the city [10]. Currently, Messina serves as a crucial center for both international and national trade, as well as for human migration, as noted in [11]. Over the past fifty years, however, the city has been grappling with challenges such as deteriorating urban conditions, a rise in youth emigration, and a shift in population towards neighboring villages, leading to a decline and spatial redistribution of its population [12]. Various factors, including economic, political and familial considerations, have attracted multiple waves of immigrants to Messina since 1930s. From the 1980s onwards, the inflows have come mostly from Sri Lanka, Philippines, Romania and Morocco.

The first four immigrants' groups settled in Messina in 2022 represent 63.57% of the total foreign population (against 76.2% in 2016, indicating an increase in the ethnic diversification of the population). Their demographic characteristics show important similarities and differences (Table 1). Between 2016 and 2022, all of the four groups registered an increase in the 25+ age classes and a reduction in the youngest ages (0–25). As regard the gender composition, Sri Lankans and Filipinos are mostly gender balanced (registering also the highest mean number in household, respectively 2.1 and 2.2 as for 2022). On the other end, Romanians are mostly females while Moroccans are generally males. The main four groups of immigrants prefer to settle in the urban core of the municipality of Messina and along the seaside, where the population density is highest

(Fig. 1). Nevertheless, Sri Lankans and Romanians seem to be more dispersed than the other groups, being located in more peripheral areas up North and in the Western inner regions.

The data on immigrants' groups come from the Population Register Office of Messina, recorded as of June 30th, 2016. This dataset encompasses Sri Lankan, Filipino, Romanian and Moroccan immigrants along with their children or nephews who were born in their respective countries of origin and obtained solely the citizenship of their parents at birth. Their residential addresses have been geocoded by querying the Google Maps Geocoding API exploiting the R "ggmap" library's functionalities [13]. All the statistical analyses and plotting were carried out with the R software [14].

In the present work, the residential segregation of the selected foreign groups is analyzed through the concept of entropy, commonly used in Information Theory. Entropy is the amount of information needed to describe a probability distribution. If two outcomes (i.e. two ethnic groups) are equally likely, there is high uncertainty about what the outcome will be and high entropy. If one outcome has a higher probability, there is less uncertainty about what the outcome will be and lower entropy. Here, the Shannon's entropy metric is computed [15]. The index considers the probabilities of the categories

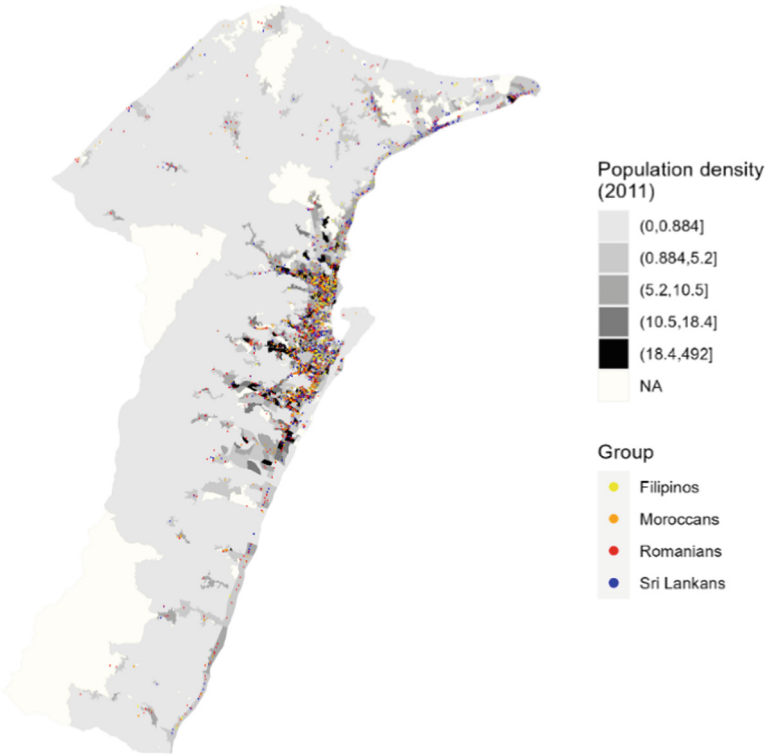


Fig. 1. Jittered foreign immigrant households' locations for the top four nationalities in Messina on June 30th, 2016 (source: Messina Population Register). In the background, population density for census tracts (source: 2011 Italian General Population Census).

Table 1. Individual characteristics of the four main immigrant ethnicities in the city of Messina at 30.06.2016 and 30.11.2022. Source: authors' elaborations on Messina Population Register.

Characteristic	Sri Lanka		Philippines		Romania		Morocco	
	2016	2022	2016	2022	2016	2022	2016	2022
Age								
0–25	30.7%	28.1%	29.7%	26.8%	21.7%	19.7%	28.8%	25.8%
26–65	66.7%	66.7%	66.2%	65.7%	76.7%	76.3%	66.6%	68.9%
65+	2.6%	5.2%	4.1%	7.5%	1.6%	4.0%	4.6%	5.3%
Sex								
Female	46.8%	46.1%	52.8%	53.7%	68.0%	68.7%	36.6%	31.6%
Male	53.2%	53.9%	47.2%	46.3%	32.0%	31.3%	63.4%	68.4%
N° of individuals	4199	7118	2555	3500	1661	2249	1218	2292
% of total immigrant population	33.2%	29.8%	20.2%	14.7%	13.1%	9.4%	9.6%	9.6%
N° of households	1742	3463	902	1609	1015	1531	573	1460
% of total number of immigrants' households	27.6%	24.1%	14.3%	11.1%	16.1%	10.6%	9.1%	10.2%
Mean n° of individuals in households	2.4	2.1	2.8	2.2	1.6	1.5	2.1	1.6

of the variable under study. Given a categorical variable X with I possible outcomes (or groups), Shannon's entropy is defined as:

$$H(X) = \sum_{i=1}^{I} p(x_i) log \frac{1}{p(x_i)} \tag{1}$$

where $p(x_i)$ for $i = 1, 2, \ldots, I$ is the probability of occurrence of group i in a given area. When populations are more mixed, $H(X)$ increases (resulting in high entropy). Conversely, when populations are more segregated, $H(X)$ decreases (yielding low entropy). Being computed over conventional areal units, such as census tracts or city blocks, the Shannon's index varies depending on the type of aggregation, the area and the uneven population counts of the preset geographical units. To overcome these flaws, the relative entropy, measuring the difference between the local population's and immigrant group's probability distributions [3], is assessed. To this end, the Kullback-Leibler (KL) divergence [16] is computed to assess the difference in the spatial arrangement between Italians and the four immigrant groups (one at the time) as:

$$D(p|q) = \sum\nolimits_{m=1}^{M} p_m log \frac{p_m}{q_m} \tag{2}$$

where the q distribution defines the standard (Italians) against which the p distribution of the immigrant group is compared in a given area m.

3 Results

In 2016, one-group Shannon's entropy measured over the six districts of Messina reveals that the largest the group, the lowest the segregation, with Sri Lankans and Filipinos being more evenly distributed across the city. The local population account for the lowest level

Table 2. Shannon's entropy index and KL divergence (against Italians) computed for the selected immigrant groups and for Italians in Messina, 2016.

Group	Shannon's entropy	KL divergence
Sri Lankans	0.129	0.070
Filipinos	0.087	0.207
Romanians	0.061	0.050
Moroccans	0.047	0.169
Italians	1.64	–

of segregation with a Shannon's index equal to 1.64. The result is not quite surprising because the level of segregation is influenced by the numerosity of the considered group. Nevertheless, the KL divergence captures a different nuance of segregation, comparing the immigrant groups' arrangement against the local population's one. In this case, Filipinos and Moroccans appear the most segregated, while Romanians' distribution diverges the least from the Italians' one across districts (Table 2).

4 Conclusions

The present study delves into the complexity of residential segregation dynamics within the metropolitan area of Messina in 2016, employing an Information Theory framework. The examination of entropy indices highlights the interplay between population distribution and segregation. While Shannon's entropy elucidates the overall level of spatial heterogeneity within the city, KL divergence unveils at what extent immigrant groups' distributions deviate from the local population's pattern. Preliminary results about residential segregation of the main immigrants' groups in 2016 show that the concentration assessed via the Shannon's metrics does not always imply a divergence between the immigrant group's and the locals' settlements. Yet, KL divergence results suggest disparities in segregation levels across immigrants' groups that interact with and are exposed to the same environments as the local population. This could imply heterogeneous levels of integration between the different immigrant groups and the local population, not just in terms of sharing the same neighborhoods, but also similar workplaces and social contexts. Spatial settlements tell us a story which reflects how population groups interact depending on group-specific cultural and social identity, capacity to afford different residential solutions, just to elucidate the main drivers of residential allocation. For this reason, tailored policies and programs fostering immigrants' integration should be implemented accounting for the spatial and contextual factors in the destination societies and the cultural and socioeconomic specificities of each immigrant group.

Acknowledgements. This study has been partially supported by the PRIN-PNRR research project "Foreign population and territory: integration processes, demographic imbalances, challenges and opportunities for the social and economic sustainability of the different local contexts (For.Pop.Ter)" [P2022WNLM7], funded by European Union—Next Generation EU, component M4C2, Investment 1.1. It has also been partially supported by the Italian Ministerial grant PRIN 2022 "Depopulation Risk: Experiences, Mobility and Subjective Well-being (DREAMS)" [2022RNKSEL] CUP E53D23010380006. It has also been partially supported by the PNRR - Missione 4 "Istruzione e Ricerca" - Componente 2 "Dalla ricerca all'impresa" - Investimento 1.3 - *Partenariato Esteso "GRINS – Growing Resilient, INclusive and Sustainable",* cod. PE0000018, CUP E63C22002120006, Spoke 8 - "*Social sustainability",* finanziato dall'Unione Europea – NextGenerationEU.

References

1. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Soc. Forces **67**, 281–315 (1988)
2. Coulter, P.B.: Measuring Inequality: A Methodological Handbook. Westview Press, Boulder (1989)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, Hoboken (2006)
4. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
5. Theil, H.: Economics and Information Theory. North-Holland Publishing Company, Amsterdam (1967)
6. Theil, H.: Statistical Decomposition Analysis. North-Holland Publishing Company, Amsterdam (1972)
7. Reardon, S.F., Firebaugh, G.: Measures of multigroup segregation. Sociol. Methodol. **32**, 33–67 (2002)
8. White, M.J.: Segregation and diversity measures in population distribution. Popul. Index **52**(2), 198–221 (1986)
9. Campione, G.: La furia di Poseidon. Messina 1908 e dintorni. Silvana Editoriale **1**, 360 (2009)
10. Bitonti, F., Mazza, A., Mucciardi, M., Scrofani, L.: Urban transformations and the spatial distribution of foreign immigrants in Messina. In: Brentari, E., Chiodi, M., Wit, E. (eds.) Models for Data Analysis Selected papers of 49th Meeting of Italian Statistical Society. Springer Proceedings in Mathematics & Statistics, pp. 53–67 (2023)
11. Casablanca, A.: Impatto ambientale e inquinamento a Messina. Armando Siciliano Editore, Messina (2001)
12. Scrofani, L.: Le aree urbane nei processi di periferizzazione e di sviluppo del Mezzogiorno. In: Sussidiarietà e… giovani al Sud Rapporto sulla Sussidiarietà 2017/2018. Fondazione Sussidiarietà, pp. 167–198 (2018)
13. Kahle, D., Wickham, H.: Ggmap: spatial visualization with ggplot2. R J. **5**(1), 144–161 (2013)
14. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2023)
15. Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
16. Kullback, S.: Letters to the editor. Am. Stat. **41**, 338–341 (1987)

Multiscale Dimensions of Residential Segregation in Naples. A Preliminary Investigation

Alessio Buonomo(✉), Federico Benassi, Rosaria Simone, and Salvatore Strozza

University of Naples Federico II, Naples, Italy
alessio.buonomo@unina.it

Abstract. The paper proposes a preliminary investigation on the multiscale dimensions of residential segregation. The study case refers to the municipality of Naples and to the most important foreign group settled in the city of Naples: the Sri Lankans. The results of the analysis, carried out at local level with the 2021 Census data and using a multiscale geographically weighted regression approach, underlines the existence of spatial heterogeneity and scale in the determinant of Sri Lankans residential segregation measured in terms of location quotient.

Keywords: residential segregation · spatial demography · multiscale geographically weighted regression

1 Introduction

Since 2008, for the first time in its history, humanity has become urban as more than half of the world's citizens reside in urban areas [1]. Cities are the main causes of wealth and progress [2] but also of strong spatial inequalities [3]. The latter usually affect the most vulnerable populations undermining social cohesion and well-being [4].

Foreign immigrants are strongly attracted by cities and, due to several factors, they are more exposed to experience spatial inequalities like residential segregation [5]. Residential segregation refers to the disproportionate distribution of population groups across a geographical area [6] and it is a process that can be seen as the opposite of the integration of the immigrants into the (urban) social fabric of the host societies [7]. High level of residential segregation and spatial inequalities of certain population groups can give

The paper has been conceived and realized as a part of the PRIN 2022-PNRR research project "Foreign population and territory: integration processes, demographic imbalances, challenges and opportunities for the social and economic sustainability of the different local contexts (For.Pop.Ter)" [P2022 WNLM7], funded by European Union-Next Generation EU, Component M4C2, Investment 1.1. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 193–198, 2025.
https://doi.org/10.1007/978-3-031-64346-0_33

rise to the so-called vicious circle of segregation [8] favoring polarization processes and the creation of dual cities [9].

Starting with the experience of the Chicago Ecological School [10], there have been many studies devoted to the subject over the years, also with reference to Europe [11–13]. With reference to the Italian context, especially in more recent years, several studies on the subject have been carried out [14, 15]. The latest studies on the topic, have shown the relevance of facing the phenomenon of residential segregation through local approaches [16–18] even if, at least for the Italian context, the multiscale dimensions of such processes have been less explored.

2 Measuring Residential Segregation

From the publication of the seminal contribution of Duncan and Duncan [19], residential segregation has progressively become a well-studied subject. Massey and Denton [20] were the first scholars to define segregation as a multidimensional concept, identifying the different dimensions that can be measured by different indices: evenness, exposure, concentration, centralization and clustering. Over the years, several indices for each dimension of segregation have been proposed by scholars as well as different approaches to classify them. The literature on the subject is vast and constantly evolving, partly as a result of technological innovations and the availability of increasingly precise and widespread georeferenced data [21].

A key issue in approaching residential segregation is its multi-scalar nature. The issue of scale is important not only in measuring and describing patterns of segregation, but also in understanding both its causes and effects [22]. This was first recognized by Duncan et al. [23] and it has been explored by many other scholars who have proposed different methodologies to face with it [24, 25]. Nevertheless, in spite of growing attention to the multi-scalar dimension of segregation, more needs to be done to examine the causes and consequences of segregation at different scales.

This paper proposes a first reflection on such topic investigating process of spatial heterogeneity and scale with reference to the level of local residential segregation of Sri Lankan citizens resident in Naples in 2021. To this aim an extension of geographically weighted regression model [26], in which the bandwidth is not assumed as constant but variable, so called multiscale geographically weighted regression (MGWR) model [27], is estimated.

3 Empirical Analysis

Data used in the paper come from the Italian Census of 2011 and 2021 and they refer to very fine geographical level (enumeration areas). A word of caution should be made on the use of such territorial units since their irregular forms can imply problems of zoning and aggregation [28]. The geographical context of study is the city of Naples where the 2021 census enumerated more than 920 thousand residents of which about 53 thousand with a foreign citizenship (5.8% of the total population).

The major foreign community is represented by Sri Lankans with slightly less than 15 thousand residents (27.8% of the total foreign population). The dependent variable

is the location quotient (LQ) [29], a local index that varies from 0 to ∞ and can detect where a particular population group (i.e., Sri Lankans in our case) is over (LQ > 1) or under-represented (LQ < 1) in comparison to the Italian population. The geographical distribution of the dependent variable shows clear spatial patterns (Fig. 1). Areas of major over representation are in the historic center of the city and in some neighbourhoods located on the seaside, where upper class reside.

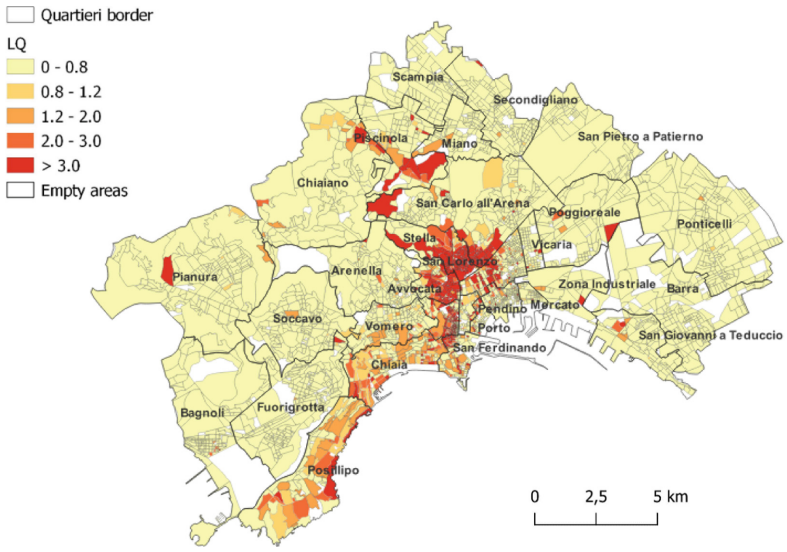


Fig. 1. Location Quotient of Sri Lankans. Municipality of Naples, 2021

Independent variables refer to three dimensions: *ethno-geographical* (x_1 = percentage of the first 9 foreign communities -excluding Sri Lankans- over the total foreigners in 2021; x_2 = percentage of foreign population over total population in 2011; x_3 = percentage of foreign women in the total foreign population in 2021), *demographical* (x_4 = mean growth rate of Italian population from 2011 to 2021; x_5 = mean growth rate of foreign population from 2011 to 2021; x_6 = average dimension of households in 2021), *labour and building environment* (x_7 = employment rate of Italian population in 2021; x_8 = percentage of building in poor condition in 2011). Two regression models have been estimated. An OLS model that serves as benchmark and a MGWR model. Assuming that there are *n* observations, for observation $i \in \{1, 2, \ldots, n\}$ at location (φ_i, δ_i), a MGWR is:

$$y_i = \beta_0(\varphi_i, \delta_i) + \Sigma_j \beta_{bwj}(\varphi_i, \delta_i) x_{ij} + \varepsilon_i \quad (1)$$

where *bwj* in β_{bwj} indicates the bandwidth used to calibrate the j^{th} conditional relationship. The idea beyond the MGWR model is that the scale of a spatial non-stationarity relationship may vary for each predictor variable. The MGWR model can differentiate local, regional, and global processes by optimizing a different bandwidth for each

covariate [30]. Variables have been standardized to allow a comparison between different bandwidths.

4 Results and Conclusions

The MGWR performs better than the OLS model having a higher R^2 (0.88 versus 0.64) and a lower AIC (1244.393 versus 2232.132). The variables are all statistically significant and they have different impact – at global level – on the dependent variable.

Results seem to indicate the existence of a process of spatial diversification of the geographical location of Sri Lankans not only regarding the Italian component but also in reference to the other major foreign communities resident in the city. In terms of process of spatial heterogeneity and scale it is important to note some key details. First, each of the independent variables acts at different territorial scale proving the existence of a multi-scalar process. Some variables have a very local scale of action while some other variables show a broader scale of effect. Second, the local betas show a spatial heterogeneity too. They are not always statistically significant and present a non-negligible level of spatial variability (Table 1).

Table 1. Regression results. OLS and MGWR[(a)]

Variables	Global beta (p-value)[(b)]	Bandwidth[(c)] (% of local beta with adj-t value (95%)[(d)])	Local beta (mean)	Local beta (Standard deviation)
x_1	–0.362***	199 (95.0)	–0.360	0.016
x_2	0.504***	44 (89.0)	0.673	0.251
x_3	–0.046**	565 (24.5)	–0.119	0.015
x_4	–0.115***	587 (95.0)	–0.124	0.027
x_5	0.255***	112 (95.0)	0.265	0.041
x_6	–0.192***	97 (95.0)	–0.172	0.020
x_7	–0.050**	214 (0.0)	–	–
x_8	0.073***	731 (95.0)	0.069	0.003

[(a)] Spatial Kernel: Adaptive bisquare; Criterion of optimal bandwidth: AICc; [(b)]These are the regression coefficients of the OLS model; [(c)] the bandwidth is determined with the number of nearest neighbourhood for each location; [(d)] based on [31]; *** p-value < 0.001, ** p-value < 0.05.

Finally, some statistical remarks are due. It is important to report that the location quotient is zero for about 66% of the total number of observations. To work with continuous data and fit a Gaussian model, the logarithm of the squared root of the nonzero LQ values was considered. After omitting units with null LQ values and missing values for covariates, the model was estimated on a set of $n = 1287$ units. To include units with $LQ = 0$ in the analysis, a jittering approach could be advocated for the null values of LQ

(sampling from the uniform distribution over (0, min(LQ), for instance)). An alternative approach to investigate spatial heterogeneity of covariates' effect on LQ would be to dichotomize its values (setting $Y = 0$ if $LQ \leq 1$, and $Y = 1$ if $LQ > 1$) and fit GWR with Binomial data. In addition, it is not possible to investigate the multiscale dimension of the phenomenon with the currently available software.

Overall, results of our preliminary investigation prove the importance and the urgency of thinking locally for statistics and society [32] and, at the same time, the intrinsic spatial nature of demographic processes [33]. The policy implications of these findings can be found in the need to establish urban observatories on foreign presence, inequalities and residential segregation. Only through this type of institution, it can be possible to effectively govern processes that operate at the local level and at various urban geographic scales, thus requiring territorially differentiated interventions.

References

1. Davis, K.: The urbanization of the human population. In: LeGates, R.T., Stout, F. (eds.) The City Reader, pp. 43–53. Routledge (2015)
2. Glaeser, E.: Triumph of the City. Our Greatest Intervention Makes us Richer, Smarter, Greener, Healthier and Happier. The Penguin Press, London (2011)
3. van Ham, M., Tammaru, T., Ubarevičienė, R., Janssen, H.: Urban Socioeconomic Segregation and Income Inequality. A Global Perspective. Springer, Cham (2021)
4. Amin, A.: Ethnicity and the multicultural city: living with diversity. Environ Plan A **34**, 959–980 (2002)
5. Dangschat, J.S.: Space matters-marginalisation and its places. Int. J. Urban Reg. Res. **33**, 835–840 (2009)
6. Bernt, T., Volkmann, A.: Residential Segregation. Encyclopedia **3**(4), 1401–1408 (2023)
7. Massey, D.S., Denton, N.A.: Spatial assimilation as a socioeconomic outcome. Am. Sociol. Rev. 94–106 (1985)
8. van Ham, M., Tammaru, T., Janssen, H: A multilevel model of vicious circles of socioeconomic segregation. In: OECD Divided Cities: Understanding Intra-urban Inequalities, pp.127–146. OECD, Paris (2018)
9. Benassi, F., Iglesias-Pascual, R.: Local-scale residential concentration and income inequalities of the main foreign-born population groups in the Spanish urban space. Reaffirming the model of a divided city. J. Ethnic Migr. Stud. **49**(3), 673–696 (2023)
10. Park, R.E., Burgess, E.W., McKenzie, R.D.: The City. University of Chicago Press, Chicago (1925)
11. Benassi, F., Bonifazi, C., Heins, F., Lipizzi, F., Strozza, S.: Comparing residential segregation of migrant populations in selected European and urban metropolitan areas. Spat. Demogr. **8**, 269–290 (2020)
12. Benassi, F., Naccarato, A., Iglesias-Pascual, R., Salvati, L., Strozza, S.: Measuring residential segregation in multi-ethnic and unequal European cities. Int. Migr. **61**(2), 341–361 (2023)
13. Benassi, F., Iglesias-Pascual, R., Salvati, L.: Residential segregation and social diversification: exploring spatial settlement patterns of foreign population in Southern European cities. Habitat Int. **1010**, 102200 (2020)
14. Strozza, S., Benassi, F., Ferrara, R., Gallo, G.: Recent demographic trends in the major Italian urban agglomerations: the role of foreigners. Spat. Demogr. **4**, 39–70 (2016)
15. Benassi, F., Crisci, M., Matthews, S.A., Rimoldi, S.M.L.: Migrants' population, residential segregation, and metropolitan spaces-insights from the Italian experience over the last 20 years. Migr. Lett. **19**(3), 287–301 (2022)

16. Bitonti, F., Benassi, F., Mazza, A., Strozza, S.: From south Asia to southern Europe: a comparative analysis of Sri Lankans' residential segregation in the main Italian cities using high-resolution data on regular lattice geographies. Genus **79**(1), 1629–1648 (2023)
17. Benassi, F., Bitonti, F., Mazza, A., Strozza, S.: Sri Lankans' residential segregation and spatial inequalities in Southern Italy: an empirical analysis using fine-scale data on regular lattice geographies. Qual. Quant. **57**(29), 1629–1648 (2023)
18. Bitonti, F., Benassi, F., Mazza, A., Strozza, S.: Framing the residential patterns of Asian communities in three Italian cities: evidence from Milan, Rome, and Naples. Soc. Sci. **19**(2) 480
19. Duncan, O.D., Duncan, B.: A methodological analysis of segregation indexes. Am. Sociol. Rev. **20**(2), 2010–2217 (1955)
20. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Soc. Forces **67**(2), 281–315 (1988)
21. Tivadar, M.: OasisR: an R package to bring some order to the world of segregation measurement. J. Stat. Softw. **89**, 1–39 (2019)
22. Reardon, S.F., et al.: The geographic scale of metropolitan racial segregation. Demography **45**(3), 489–514 (2000)
23. Duncan, O.D., Cuzzort, R.P., Duncan, B.: Statistical Geography: Problems in Analysing Areal Data. The Free Press of Glencoe, New York (1961)
24. Fowler, C.S., Lee, B.A., Matthews, S.A.: The contributions of places to metropolitan ethno-racial diversity and segregation: decomposing change across space and time. Demography **53**(6), 1955–1977 (2016)
25. Lichter, D., Parisi, D., Ambinakudige, S., Scott, C.K.: Reevaluating the spatial scale of residential segregation: racial change within and between neighborhoods. Demography 11195639 (2024)
26. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationship. Wiley, Hoboken (2003)
27. Oshan, T., M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S.: mgwr: a Python implementation of multiscale geographically weighted regression for investigating process of spatial heterogeneity and scale. ISPRS Int. J. Geo-Inf. **8**(6), 269 (2019)
28. Openshaw, S., Taylor, P.J.: A million or so correlation coefficients: three experiments on the Modifiable Area Unit Problem. In: Wrigley, N. (ed.) Statistical Applications in the Spatial Sciences, pp. 127–144. Pion, London (1979)
29. Haig, R.M.: Toward and understanding of the metropolis. Quart. J. Econ. **40**, 421–433 (1926)
30. Li, Z., Fotheringham, A.S.: Computational improvements to multi-scale geographically weighted regression. Int. J. Geogr. Inf. Sci. **34**(7), 1378–1397 (2020)
31. Yu, H., Fotheringham, A.S., Li, Z., Oshan, T., Kang, W., Wolf, L.J.: Inference in multiscale geographically weighted regression. Geogr. Anal. **52**(1), 87–106 (2020)
32. Fotheringham, A.S., Sachdeva, M.: On the importance thinking locally for statistics and society. Spat. Stat. **50**, 100601 (2022)
33. Voss, P.R.: Demography as a spatial social science. Popul. Res. Policy Rev. Policy Rev. **26**, 457–476 (2007)

Socioeconomic Distress and Foreign Presence in a Southern Urban Context. The Case of Bari

Maria Carella(✉), Thaís García-Pereiro, and Anna Paterno

University of Bari "Aldo Moro", Bari, Italy
maria.carella1@uniba.it

Abstract. Recent studies found a clear North-South hierarchy of urban areas in Europe within the context of growing multiculturalism and socioeconomic inequality. The objectives of this article are three-fold. The first is the identification of hotspots of spatial inequalities after assessing multiple socioeconomic distress. The second is to examine the spatial distribution of non-national population through their overrepresentation respect to Italian nationals. The last is to measure the local spatial correlation between the degree of potential socioeconomic distress and the concentration of the non-nationals' population.

Keywords: Socioeconomic distress · foreign population · spatial analysis · Bari

1 Introduction

Recent research shows that social segregation in European cities increased due to the substantial and consistent growth of immigration flows and the rise of socioeconomic inequalities (Andersen and van Kempen 2003, Lymperopoulou and Finney 2017).

Undoubtedly, socioeconomic and housing conditions of the urban placement and its surrounding places play a relevant role not only on the size of the foreign population settling there, but also on its spatial patterns of residential segregation (Marcińczak et al. 2021, Pisarevskaya et al. 2021). This is particularly true when understanding residential segregation as the extent to which individuals from different groups (socioeconomic status, ethnicity, etc.) inhabit and actively live different locations (Reardon and O'Sullivan 2004).

The latest studies on this subject find a clear North-South hierarchy of urban areas in Europe within the context of growing multiculturalism and socioeconomic inequality. That is, southern urban areas holding higher levels of segregation are also those combining a weaker economy with higher degrees of social vulnerability (Benassi et al. 2020, Marcińczak et al. 2021, Benassi et al. 2022).

The objectives of this article are threefold. It first presents an approach to assess multiple socioeconomic distress across urban populations at a local scale (census tracts). This method is applied to the city of Bari, Apulia, to illustrate its usability for identifying hotspots of spatial distress. Secondly, it examines the spatial distribution of non-national population through their overrepresentation respect to Italian nationals, to give a clearer idea of the suburban geography of migrant groups in the city. And, finally, it measures

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 199–204, 2025.
https://doi.org/10.1007/978-3-031-64346-0_34

the local spatial correlation between the degree of potential socioeconomic distress and the concentration of the non-nationals' population. Results might help stakeholders and policymakers to screen the extent of potential distress locally and across population subgroups and support knowledge-based policies, in particular regeneration policies in disadvantaged neighbourhoods.

2 Main Theories Briefly Explained

Contemporary cities are becoming one of the most significant mechanisms of selection, stratification and even socioeconomic exclusion (Andersen and Van Kempen 2001). The transformation of urban agglomerations produced has progressively reconfigured public areas and private living spaces (Montalbano 2020, Piketty 2020), leading to different spatial patterns of population in the cities. Moreover, the competition to occupy residential space has often result in profound spatial segmentation of urban contexts (Van Kempen, 2007) revealing internal socio-spatial inequalities (Yao et al. 2019).

The concept of the dual city has been used to highlight the socio-economic inequalities in the cities (Castells and Mollenkopf 1991) and polarized urban spaces (Fainstein 1992). A strand of literature has framed the division of urban space within the context of socio-spatial segregation of large cities (Oberti and Préteceille 2016), often measuring the differences between neighborhoods according to the resources of their residents (Maloutas and Spyrellis 2019). This phenomenon has been investigated by different approaches mainly based on the ethno-racial differences in the occupation of urban space (Benassi et al. 2020) and on the perspective of social classes within large urban areas (Oberti and Préteceille 2004).

Several studies have expanded the dual city idea enriching the meaning of its socio-spatial patterning (Benassi and Iglesias Pascual 2023). But what about the intermediate situations between the two extreme poles? How can we bring into this interpretative scheme 'the multiple/plural city' or the multiple dual cities composed of a myriad of neighbourhoods with different economic, social, environmental and political situations regardless of their internal inequalities?

Currently, especially in Southern Europe, economic uncertainty is progressively reducing income of families. As a result, the most fragile ones (single-parent families, multigenerational families, etc.) are often moving out of the large conurbations (Palomares-Linares 2020) and the urban space suffers a breakup within it. Some studies have documented that neighbourhoods in many southern cities are characterised by a variety of situations that affect individuals producing a strong division of the inner parts of the contemporary city (Arbaci 2019, Tammaru et al. 2020). Thus, the city becomes a mosaic of multiple spaces in which neighbourhoods or districts hit by vulnerability or urban distress coexist with wealthier contexts (Paquot 2002).

Based on the former, this article is aimed at answering the following research questions:

RQ1. Are there any differences in the suburban distribution of the Index of Potential Socioeconomic Distress in the city?
RQ2. Are there important contrasts in the residential and concentration patterns of the population of non-nationals and nationals across the city?

RQ3. Is there a spatial relationship between suburban patterns of potential socioeconomic distress and territorial concentration of non-nationals respect to nationals?

3 Data and Methodology

Our empirical analyses are based on the last available Census data for 2021 at the level of the census tracts of the city of Bari from the information publicly provided by the Italian National Statistics Institute (ISTAT). More specifically, we use data on the age, nationality, number of household members, level of education, non-employment status of resident population, and housing conditions[1]. We select census tracts having at least 10 residents (n = 1,291).

To measure potential socioeconomic distress, we built a composite indicator based on six normalized items covering three different dimensions: sociodemographic (shares of individuals over 70 among total population and households with more than four components among total households), socioeconomic (low human capital by gender: shares of male/female population with at most the first level of secondary education among total males/females; and non-employment by gender: not-employed males/females among total male/female population 15–64) and housing conditions (share of residential buildings in bad or very bad state of preservation among total residential buildings). By applying Principal Component Analysis (PCA) we reduced the number of indicators considered to a smaller number of principal components (3)[2] accounting for most of the observed variation (73% of data total variance). The composite index was built up for each census tract with the PCs retained, weighted by their eigenvalues. Finally, the indicator was normalized to obtain a Composite Index of Potential Socioeconomic Distress that varies between 0 (null distress) and 1 (maximum distress). The values of the index are represented in Fig. 1 at the level of census tracts using a natural breaks (Jenks) map.

For the analysis of the spatial concentration patterns of the non-national population, we compute a Local Quotient (LQ) as a ratio of ratios (Benassi and Iglesias-Pascual 2023). In the first ratio the total population of non-nationals is divided among the total Italian population for the whole city; the second ratio is the same, but it is computed for each census tract. Then, if LQ is minor than 1(LQ $<$ 1) non-nationals are underrepresented respect to nationals, instead, if LQ is greater than 1(LQ $>$ 1) non-nationals are over-represented.

In the last step we estimate both the global and local version of bivariate Moran's I between the Composite Index of Potential Socioeconomic Distress and the obtained LQs. For the sake of briefness, we are not able to show these results here, but they are available upon request.

[1] Information on the state of buildings was drawn from the 2011 Census because it was not available for 2021. This might not affect results given that the state of buildings are structural conditions that tend to remain stable over time.

[2] For the sake of brevity, results of PCA are not shown here but available upon request.

4 Selected Results

As shown by the natural breaks map (Fig. 1), there are relevant differences in the suburban distribution of the Composite Index of Potential Socioeconomic Distress in the city of Bari. Greener spots are those showing lower levels of distress, while red ones indicate those areas more potentially distressed. There is an "island-type" spatial distribution in both green and red spots. Concentrating the attention in the red spots, it seems that the sub-urban areas of high distress are not only spatially clustered but also clustered-disperse, which might be indicating its duality-in terms of spatial isolation patterns of the most vulnerable groups in certain areas of the city.

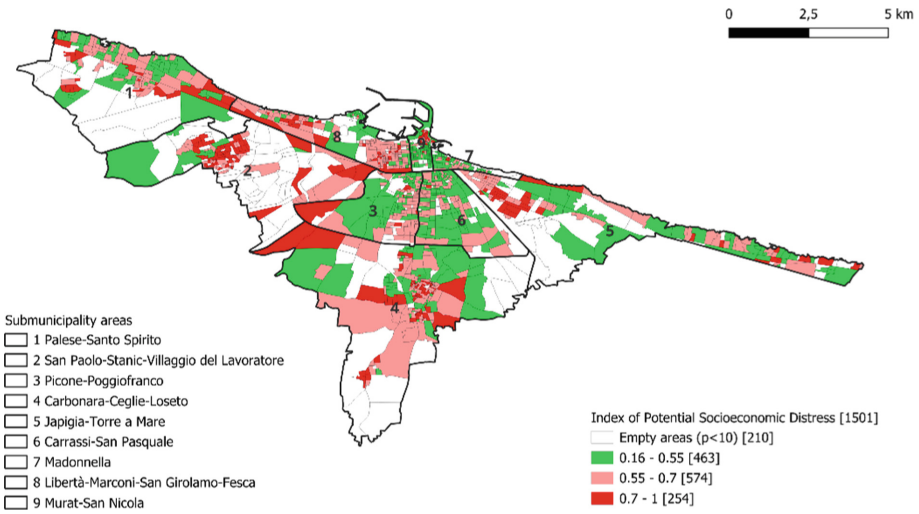


Fig. 1. Composite Index of Potential Socioeconomic Distress. Bari (2021). Classes obtained by natural breaks (Jenks) method. In 2014, sub-municipality areas were substituted by 5 municipalities. We believe that the first are more informative (less concentrated) and use them instead of the latter.

Figure 2 illustrates the LQs according to which the population groups of non-nationals display notably dissimilar spatial distributions respect to the native population. It deserves to be highlighted that the larger number census tracts in which non-nationals are over-represented if compared to nationals are those of very high over-representation ($n = 243$). A first glance to combine information of both figures allow us to note that suburban areas where the potential distress is higher are not necessarily those in which non-nationals are over-represented. In fact, the results from the local bivariate Moran's I index are indicating important disparities. Similar values of potential distress and non-nationals' over-representation tend to diverge in the space (with high-low clusters prevailing over high-high and low-low clusters).

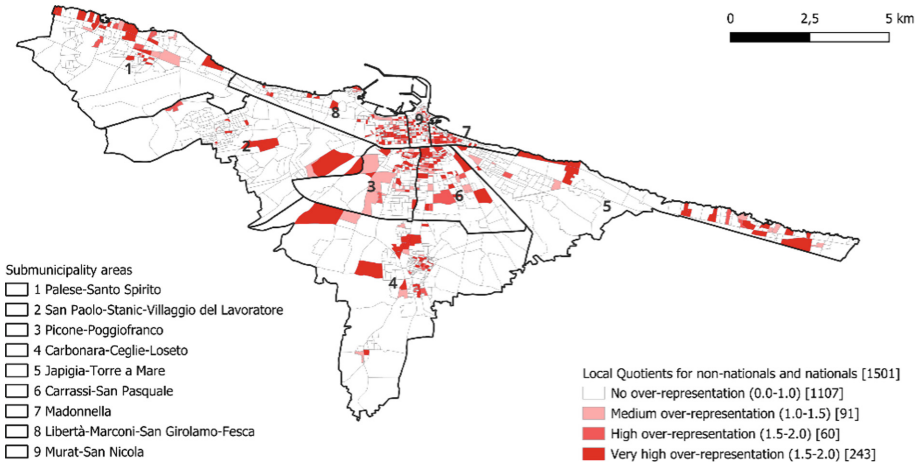


Fig. 2. Local Quotients (LQ) for non-national and national population subgroups. Bari (2021).

Acknowledgments. This study was conceived and realized as part of the PRIN-PNRR research project "Foreign population and territory: integration processes, demographic imbalances, challenges and opportunities for the social and economic sustainability of the different local contexts (For.Pop.Ter)" [P2022 WNLM7], funded by European Union—Next Generation EU, component M4C2, Investment 1.1. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission or the authors' own institutions. Neither the European Union nor the European Commission can be held responsible for them.

References

Andersen, H.T., Van Kempen, R.: Governing European Cities: Social Fragmentation, Social Exclusion and Urban Governance. Aldershot, Ashgate (2001)

Andersen, H.T., Van Kempen, R.: New trends in urban policies in Europe: evidence from the Netherlands and Denmark. Cities **20**(2), 77–86 (2003)

Arbaci, S.: Paradoxes of Segregation: Housing Systems, Welfare Regimes and Ethnic Residential Change in Southern European Cities. Wiley, Hoboken (2019)

Benassi, F., Iglesias-Pascual, R.: Local-scale residential concentration and income inequalities of the main foreign-born population groups in the Spanish urban space. Reaffirming the model of a divided city. J. Ethnic Migr. Stud. **49**(3), 673–696 (2023). https://doi.org/10.1080/1369183X.2022.2067137

Benassi, F., Naccarato, A., Iglesias-Pascual, R., Salvati, L., Strozza, S.: Measuring residential segregation in multi-ethnic and unequal European cities. Int. Migr.Migr. **61**(2), 341–361 (2022)

Benassi, F., Bonifazi, C., Heins, F., Lipizzi, F., Strozza, S.: Comparing residential segregation of migrant populations in selected European urban and metropolitan areas. Spat. Demography **8**, 269–290 (2020). https://doi.org/10.1007/s4098

Castells, M., Mollenkopf, J.H.: Dual City: Restructuring New York. Russell Sage Foundation, New York (1991)

Fainstein, S., Gordon, I., Harloe, M., Divided cities, Cambridge, Blackwell 1992

Lymperopoulou, K., Finney, N.: Socio-spatial factors associated with ethnic inequalities in districts of England and Wales, 2001–2011. Urban Stud. **54**(11), 2540–2560 (2017). https://doi.org/10.1177/0042098016653725

Marcińczak, S., Mooses, V., Strömgren, M., Tammaru, T.: A comparative study of immigrant-native segregation at multiple spatial scales in urban Europe. J. Ethnic Migr. Stud. 1–23 (2021). https://doi.org/10.1080/1369183X.2021.2008887

Maloutas, T., Spyrellis, S.N.: Segregation trends in Athens: the changing residential distribution of occupational categories during the 2000s. Reg. Stud. **54**(4), 462–471 (2019). https://doi.org/10.1080/00343404.2018.1556392

Montalbano, C.: The future of the city: the fragment and the sense of place. In: Kong, M.S.M., Monteiro, M.R., Neto, M.J.P., Gale, A.M.M., Xavier, J.P., Dias J.C. (eds.) Tradition and Innovation. CRC Press/Balkema Book/Taylor & Francis, London (2020)

Oberti, M., Préteceille, E.: La ségrégation urbaine, La Découverte, col. Repères Sociologie, p. 124 (2016)

Oberti, M., Préteceille, E.: Les classes moyennes et la ségrégation urbaine. Éducation et sociétés **14**(2), 135–153 (2004). https://doi.org/10.3917/es.014.0135

Palomares-Linares, I., Baldán, H., Torrado, J.M., Susino, J.: Making place for urban segregation matters in four southern European countries: a literature review. In: Social Problems in Southern Europe (2020)

Paquot, T.: Ville fragmentée ou urbain éparpillé? Dans Françoise Navez-Bouchanine(dir.) La fragmentation en question: des villes entre fragmentation spatiale et fragmentation sociale?, pp. 113–118. L'Harmattan, Paris (2002)

Piketty, T.: Capital and Ideology. Harvard University Press, Cambridge, MA and London, England (2020). https://doi.org/10.4159/9780674245075

Pisarevskaya, A., Scholten, P., Kaşlı, Z.: Classifying the diversity of urban diversities: an inductive analysis of European cities. J. Int. Migr. Integr.Migr. Integr. (2021). https://doi.org/10.1007/s12134-021-00851-z

Yao, J., Wong, D.W., Bailey, N., Minton, J.: Spatial segregation measures: a methodological review. Tijdschr. Econ. Soc. Geogr.. Econ. Soc. Geogr. **110**, 235–250 (2019). https://doi.org/10.1111/tesg.12305

Reardon, S.F., O'Sullivan, D.: Measures of spatial segregation. Sociol. Methodol.. Methodol. **34**, 121–162 (2004). https://doi.org/10.1111/j.0081-1750.2004.00150.x

van Kempen, R.: Divided cities in the 21st century: challenging the importance of globalization. J. Hous. Built Environ. **22**(1), 13–31 (2007)

Tammaru, T., Marcińczak, S., Aunap, R., van Ham, M., Janssen, H.: Relationship between income inequality and residential segregation of socioeconomic groups. Reg. Stud. **54**(4), 450–461 (2020). https://doi.org/10.1080/00343404.2018.1540035

van Ham, M., Tammaru, T., Ubarevičienė, R., Janssen, H.: Urban Socio-Economic Segregation and Income Inequality: A Global Perspective. The Urban Book Series. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-64569-4

Point Process Learning: A Cross-Validation-Based Statistical Framework for Point Processes

Julia Jansson[1], Christophe A. N. Biscio[2], Mehdi Moradi[3], and Ottmar Cronie[1(✉)]

[1] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
ottmar@chalmers.se

[2] Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

[3] Department of Mathematics and Mathematical Statistics, Umeå University,Umeå, Sweden

Abstract. Recently, Point Process Learning was introduced as a powerful approach to fitting Papangelou conditional intensity models to point pattern data. This cross-validation-based statistical theory was shown to significantly outperform the state-of-the-art in the context of kernel intensity estimation. In this paper, we further illustrate its potential by showing that it outperforms the state-of-the-art when fitting a hard-core Gibbs model.

Keywords: Cross-Validation · Gibbs processes · Point Process Learning · Prediction errors · Spatial statistics

1 Introduction

The general intractability of density/likelihood functions for point processes has driven the statistical development away from likelihood estimation toward methods exploiting (Papangelou) conditional intensities, which are given as density function ratios [1,4]. The most prominent conditional intensity-based approach is arguably pseudolikelihood estimation, which is obtained by replacing the intensity function in the Poisson process likelihood with the target model's conditional intensity. For different models, most notably those with strong spatial interactions, this approach may result in poor estimation performances [1].

Motivated by cross-validation procedure's general ability to reduce mean square errors in different settings, [3] introduced a novel cross-validation-based statistical theory called Point Process Learning, which has shown great promise in conditional intensity modelling. It is based on the combination of cross-validation for point processes and a notion of prediction errors for point processes.

In this paper, we explore the performance of Point Process Learning in conditional intensity modelling for a particular Gibbs process, the hard-core

O. Cronie—Supported by the Swedish Research Council (2023-03320).

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 205–211, 2025.
https://doi.org/10.1007/978-3-031-64346-0_35

process [1]. More specifically, we study some basic theoretical properties and, through a simulation study, we show that it significantly outperforms pseudo-likelihood estimation in terms of mean square error.

2 Preliminaries

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $X = \{x_i\}_{i=1}^{N}$, $N \geq 0$, be a (simple) point process in a general (Polish) space S with reference measure $A \mapsto |A| = \int_A \mathrm{d}u$, $A \subseteq S$, most notably a Euclidean domain $S \subseteq \mathbb{R}^d$, $d \geq 1$, with Lebesgue measure. Throughout, we use the notation "$\subseteq$" for Borel sets. Formally, X is a random variable [4] in the measurable space $(\mathcal{X}, \mathcal{N})$ of point patterns $\mathbf{x} = \{x_1, \ldots, x_n\} \subseteq S$, $n \geq 0$, where the cardinalities $\#(\mathbf{x} \cap A)$ are finite for bounded $A \subseteq S$. Throughout, we assume that all functions considered are sufficiently measurable/integrable.

The conditional intensity λ_X completely characterises the distribution of X. Heuristically, $\lambda_X(u; \mathbf{x})\mathrm{d}u$ gives the probability of X having a point in an infinitesimal neighbourhood du of $u \in S$, conditionally of X agreeing with $\mathbf{x} \in \mathcal{X}$ outside du [4]. Note in particular that the intensity function of X is given by $\rho_X(u) = \mathbb{E}[\lambda_X(u; X)]$, $u \in S$. Moreover, λ_X is called repulsive (attractive) if $\lambda_X(u; \mathbf{x}) \leq \lambda_X(u; \mathbf{y})$ $(\lambda_X(u; \mathbf{x}) \geq \lambda_X(u; \mathbf{y}))$, $u \in S$, whenever $\mathbf{x} \subseteq \mathbf{y}$ [4].

The statistical problem we are dealing with here is to obtain an estimate $\widehat{\theta}$ of $\theta_0 \in \Theta$, based on a single point pattern $\mathbf{x} \in \mathcal{X}$ and a parametrised conditional intensity λ_θ, $\theta \in \Theta$, where $\lambda_X = \lambda_{\theta_0}$. The main motivation for using λ_θ, $\theta \in \Theta$, to estimate θ_0 comes from the general intractability of the associated density/likelihood function $f_\theta(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, $\theta \in \Theta$ [4].

3 Point Process Learning

The Point Process Learning approach is based on the following two concepts, which were introduced in [3]: i) point process cross-validation, and ii) point process prediction errors.

The definition of thinning for point processes through bivariate marking was formalised in [3]. More specifically, given a (potentially random) marking function $M : S \to \{0, 1\}$, an M-thinning of X is given by $X^V = \{x \in X : M(x) = 1\}$. When M acts independently on X, we call X^V an independent thinning. When, conditionally on X, the random variables $M(x)$, $x \in X$, are iid Bernoulli random variables with parameter $p \in (0, 1)$, we call X^V a p-thinning [2]. Given thinnings $X_1^V, \ldots, X_k^V$, $k \geq 1$, we call (X_i^T, X_i^V), $X_i^T = X \setminus X_i^V$, $i = 1, \ldots, k$, the training-validation pairs of a cross-validation (CV). When all training-validation pairs are independently generated p-thinnings, we speak of Monte-Carlo CV (MCCV); see [3] for further alternatives. For a point pattern $\mathbf{x}$, all definitions are analogous, and we denote a CV round by $\{(\mathbf{x}_i^T, \mathbf{x}_i^V)\}_{i=1}^{k}$.

Consider a training-validation pair (X^T, X^V), a model λ_θ, $\theta \in \Theta$, where λ_{θ_0} corresponds to $X = X^T \cup X^V$, and a so-called test function $\mathcal{H}_\Theta = \{h_\theta : \theta \in \Theta\}$, $h_\theta : S \times \mathcal{X} \to \mathbb{R}$. The associated $\mathcal{H}_\Theta$-weighted prediction errors are given by

$$\mathcal{I}^{h_\theta}_{\lambda_\theta}(A; X^T, X^V) = \sum_{x \in X^V \cap A} h_\theta(x; X^T) - \int_A h_\theta(u; X^T) w(u) \lambda_\theta(u; X^T) \mathrm{d}u \quad (1)$$

for $A \subseteq S$, where $w(u)$, $u \in S$, is a specific (random) weight function. Here, the sum term embodies the "empirical" prediction of the points of X^V by X^T, through the test function. The integral term, on the other hand, is essentially a model description of the sum term.

The key in Point Process Learning is given in Theorem 2 in [3]. Among other things, it states that the expectation of $\mathcal{I}^{h_\theta}_{\lambda_\theta}(A; X^T, X^V)$ is 0 for any test function and $A \subseteq S$ if and only if the weight function $w(u)$, $u \in S$, has a particular form and $\theta = \theta_0$. The proper weight function is given as a conditional expectation, which is not necessarily tractable. However, Corollary 1 in [3] tells us that e.g. under MCCV we have that i) $w(\cdot) \in (0, p]$ if we consider a repulsive model, ii) $w(\cdot) \geq p$ if we consider an attractive model, and iii) $w(\cdot) = p$ if we have a Poisson process model. The general observation in [3] is that for both repulsive and Poisson process models, it makes sense to simply set $w(\cdot) = p$, while for attractive models $w(\cdot) = p/(1-p)$ seems to be a sensible choice. Motivated by these observations, [3] introduced the following loss functions to be minimised:

$$\begin{aligned}
\mathcal{L}_j(\theta) =& \frac{1}{\#\mathcal{T}_k} \sum_{i \in \mathcal{T}_k} |\mathcal{I}^{h_\theta}_{\lambda_\theta}(W; \mathbf{x}_i^V, \mathbf{x}_i^T)|^j, \quad j = 1, 2, \\
\mathcal{L}_3(\theta) =& \left(\frac{1}{\#\mathcal{T}_k} \sum_{i \in \mathcal{T}_k} \mathcal{I}^{h_\theta}_{\lambda_\theta}(W; \mathbf{x}_i^V, \mathbf{x}_i^T) \right)^2, \\
\mathcal{T}_k =& \{i \in \{1, \ldots, k\} : 1 \leq \#\mathbf{x}_i^T \leq \#\mathbf{x} - 1\}.
\end{aligned} \quad (2)$$

It typically makes no sense to predict $\mathbf{x}_i^V = \mathbf{x}$ from $\mathbf{x}_i^T = \emptyset$, or to predict $\mathbf{x}_i^V = \emptyset$ from $\mathbf{x}_i^T = \mathbf{x}$, which is why the sums run over $\mathcal{T}_k$. We are, of course, free to let $\mathcal{T}_k = \{1, \ldots, k\}$ if it makes sense in a given setting.

4 Hard-Core Process Modelling

We here study Point Process Learning in the context of hard-core processes, with conditional intensities $\lambda_\theta(u; \mathbf{x}) = \beta_{\theta'}(u) \mathbf{1}\left\{u \notin \bigcup_{x \in \mathbf{x}} b(x, R)\right\}$, $u \in S$, $\mathbf{x} \in \mathcal{X}$, where the functions $\beta_{\theta'}(\cdot)$, $\theta' \in \Theta' \subseteq \mathbb{R}^{l'}$, $l' \geq 1$, are arbitrary and $\theta = (\theta', R) \in \Theta = \Theta' \times (0, \infty)$. The true parameters of X are denoted by $\theta_0 = (\theta_0', R_0)$.

4.1 State-of-the-Art

The likelihood estimate of R_0 is given by [4, Example 3.17]

$$\bar{R} = \min_{x, y \in \mathbf{x}, x \neq y} d(x, y),$$

but to the best of our knowledge, a closed form likelihood estimator for θ_0' is not available in the literature [4]. Turning to pseudolikelihood estimation [1], where

we maximise $PL(\theta) = \sum_{x\in\mathbf{x}} \log \lambda_\theta(x; \mathbf{x}\setminus\{x\}) - \int_S \lambda_\theta(u;\mathbf{x})\mathrm{d}u$, $\theta \in \Theta$, with the convention that $\log 0 = -\infty$, it can be shown here that the pseudolikelihood estimate $\widehat{R}_{PL}$ of R_0 must belong to $(0, \bar{R})$; for a fixed θ', the function PL is constant for $R \in (0, \bar{R})$, whereby we have identifiability issues. To carry out the estimation of θ', we next set the gradient of the resulting expression to 0, i.e.

$$\sum_{x\in\mathbf{x}} \frac{\nabla_{\theta'}\beta_{\theta'}(x)}{\beta_{\theta'}(x)} - \int_{S\setminus\bigcup_{x\in\mathbf{x}} b(x,R)} \nabla_{\theta'}\beta_{\theta'}(u)\mathrm{d}u = 0 \in \mathbb{R}^{l'}, \quad R \in (0, \bar{R}),$$

where we assume that $\beta_{\theta'}(\cdot)$ is such that integration and differentiation may be interchanged. In particular, when $\beta_{\theta'}(\cdot)$ is constant, i.e. $\beta_{\theta'}(\cdot) \equiv \beta \in \Theta' = (0, \infty)$, we obtain

$$\widehat{\beta}_{PL} = \widehat{\beta}_{PL}(\widehat{R}_{PL}) = \frac{|S|}{|S\setminus\bigcup_{x\in\mathbf{x}} b(x, \widehat{R}_{PL})|}\widetilde{\theta}(\mathbf{x}, S), \quad \widehat{R}_{PL} \in (0, \bar{R}),$$

an adjusted version of the classical homogeneous intensity estimate $\widetilde{\theta}(\mathbf{x}, S) = \#\mathbf{x}/|S|$. Note that $\widehat{\beta}_{PL}(\widehat{R}_{PL})$ decreases as $\widehat{R}_{PL}$ increases. Practically, when $S \subseteq \mathbb{R}^2$, this estimation may be done with the function `ppm` in the R package `spatstat` [1], which uses the choice $\widehat{R}_{PL} = \bar{R}\#\mathbf{x}/(\#\mathbf{x} + 1)$.

4.2 Point Process Learning

Turning to the Point Process Learning approach, we here consider p-thinning CV (e.g. MCCV), the weight function $w(\cdot) = p$ and a test function $h_\theta(u;\mathbf{x}) = f(p\lambda_\theta(u;\mathbf{x}))$, where $f : \mathbb{R} \to \mathbb{R}$ is such that $\lim_{x\to 0} |f(x)| = \infty$. This includes e.g. $f(x) = x^{-\gamma}$, $\gamma > 0$ [3]. Given $\mathbf{x}^T, \mathbf{x}^V \neq \emptyset$, the prediction errors in (1) become

$$\begin{aligned}\mathcal{I}^{h_\theta}_{\lambda_\theta}(S; \mathbf{x}^V, \mathbf{x}^T) = &\sum_{x\in\mathbf{x}^V} f(p\beta(x)\mathbf{1}\{x \notin \bigcup_{y\in\mathbf{x}^T} b(y, R)\}) \\ &- p\int_{S\setminus\bigcup_{x\in\mathbf{x}^T} b(x,R)} \beta(u) f(p\beta(u)\mathbf{1}\{u \notin \bigcup_{x\in\mathbf{x}^T} b(x, R)\})\mathrm{d}u.\end{aligned}$$

It follows that the loss functions are finite only if R belongs to

$$\mathcal{R} = \mathcal{R}_p(\{(\mathbf{x}_i^V, \mathbf{x}_i^T)\}_{i=1}^k) = \left\{ r > 0 : \mathbf{x}_i^V \cap \bigcup_{x\in\mathbf{x}_i^T} b(x, r) = \emptyset \text{ for all } i \in \mathcal{T}_k \right\},$$

where we recall $\mathcal{T}_k$ from (2). In other words, the estimate $\widehat{R} = \widehat{R}_p(\{(\mathbf{x}_i^V, \mathbf{x}_i^T)\}_{i=1}^k)$ of the interaction/hard-core range R_0 belongs to $\mathcal{R}$ and in the MCCV case we obtain that $\lim_{k\to\infty} \mathcal{R}_p(\{(\mathbf{x}_i^V, \mathbf{x}_i^T)\}_{i=1}^k) = (0, \bar{R})$. By imposing the restriction that $\theta = (\theta', R) \in \Theta' \times \mathcal{R}$, the prediction errors reduce to

$$\sum_{x\in\mathbf{x}_i^V} f(p\beta_{\theta'}(x)) - p\int_{S\setminus\bigcup_{x\in\mathbf{x}_i^T} b(x,R)} f(p\beta_{\theta'}(u))\beta_{\theta'}(u)\mathrm{d}u, \tag{3}$$

i.e. the loss function for estimating θ_0' is given by a combination of the terms in (3), for $\theta = (\theta', R) \in \Theta' \times \mathcal{R}$ and $i \in \mathcal{T}_k$. Since $\mathcal{I}^{h_{\theta_1}}_{\lambda_{\theta_1}}(S; \mathbf{x}_i^V, \mathbf{x}_i^T) = \mathcal{I}^{h_{\theta_2}}_{\lambda_{\theta_2}}(S; \mathbf{x}_i^V, \mathbf{x}_i^T)$ does not imply that $\theta_1 = \theta_2$ (if $\theta_1 = (\theta', R_1)$ and $\theta_2 = (\theta', R_2)$, these two prediction errors are the same for any $R_1, R_2 \in \mathcal{R}$), the loss function $\theta \mapsto \mathcal{I}^{h_\theta}_{\lambda_\theta}(S; \mathbf{x}_i^V, \mathbf{x}_i^T)$ is not identifiable for a fixed $i \in \mathcal{T}_k$. One would typically deal with this by fixing a point estimate $\widehat{R}$ of R_0, e.g. $\widehat{R} = \widehat{R}_p(\{(\mathbf{x}_i^V, \mathbf{x}_i^T)\}_{i=1}^k) = \sup \mathcal{R}$, and then proceed by exploiting the prediction errors in (3) for the estimation of θ'. However, we have seen that, numerically, this is not necessary when employing any of the loss functions $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$, i.e. we may let both $R \in \mathcal{R}$ and $\theta' \in \Theta'$ be free parameters to be estimated. In other words, the component-wise unidentifiability seems to not spill over on the loss functions.

We next turn to the special case where $\beta_{\theta'}(\cdot) \equiv \beta \in \Theta' = (0, \infty)$ is constant and $\beta_{\theta_0'}(\cdot) = \beta_0 \in \Theta'$, which yields

$$\mathcal{I}^{h_\theta}_{\lambda_\theta}(S; \mathbf{x}_i^V, \mathbf{x}_i^T) = f(p\beta)\Big(\#\mathbf{x}_i^V - p\beta\Big|S \setminus \bigcup_{x \in \mathbf{x}_i^T} b(x, R)\Big|\Big), \quad i \in \mathcal{T}_k, R \in \mathcal{R}. \tag{4}$$

If we impose that $|f(x)| > 0$, $x > 0$, which e.g. holds for $f(x) = x^{-\gamma}$, $\gamma > 0$, then (4) is 0 if β is set to

$$\widehat{\beta}_i(R) = \frac{\#\mathbf{x}_i^V}{p|S \setminus \bigcup_{x \in \mathbf{x}_i^T} b(x, R)|} = \frac{|S|}{|S \setminus \bigcup_{x \in \mathbf{x}_i^T} b(x, R)|} \frac{\widetilde{\theta}(\mathbf{x}_i^V, S)}{p}, \quad R \in \mathcal{R},$$

which essentially is equivalent to a CV-based version of $\widehat{\beta}_{PL}(R)$, $R \in \mathcal{R}$. It should be emphasised that $|S \setminus \bigcup_{x \in \mathbf{x}_i^T} b(x, R)|$ is not linear in p (from a distributional point of view) so we expect the choice of p to be of significance here.

4.3 Numerical Evaluation

We next evaluate our approach numerically in the case where $\beta_{\theta'}(\cdot) \equiv \beta \in \Theta' = (0, \infty)$ and $\beta_{\theta_0'}(\cdot) = \beta_0 \in \Theta'$. More specifically, we consider 500 realisations of a hard-core model on $S = [0, 1]^2$ with $\theta_0 = (R_0, \beta_0) = (0.05, 100)$; this particular parameter choice gives rise to an average point count of 58.51.

We here consider the loss functions $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$, in combination with (4), with $\mathcal{T}_k$ as in (2), and MCCV, where $k = 100$ and $p = 0.1, 0.15, \ldots, 0.9, 0.95$. In addition, we use the test function $h_\theta(u; \mathbf{x}) = 1/(p\lambda_\theta(u; \mathbf{x}))$. To compare with the state-of-the-art, we additionally carry out pseudolikelihood estimation. In Fig. 1 we report estimates of the mean square errors $\mathrm{MSE}(\widehat{\beta}) = \mathbb{E}[(\widehat{\beta} - \beta_0)^2]$ and $\mathrm{MSE}(\widehat{R}) = \mathbb{E}[(\widehat{R} - R_0)^2]$, for each combination of loss function and p.

As anticipated and as can be observed in Fig. 1, the choice of p does have a significant impact on the performance. To begin with, we see that both $\mathcal{L}_1$, $\mathcal{L}_2$ and the pseudolikelihood estimator essentially yield perfect estimates of R_0, while $\mathcal{L}_3$ performs more poorly, as a consequence of both a higher bias and a higher variance. In the case of β_0, where the performance for the pseudolikelihood estimator is $\mathrm{MSE}(\widehat{\beta}) = 314.8607$, we find that $\mathcal{L}_1$ and $\mathcal{L}_2$ with $p \leq 0.3$ have the

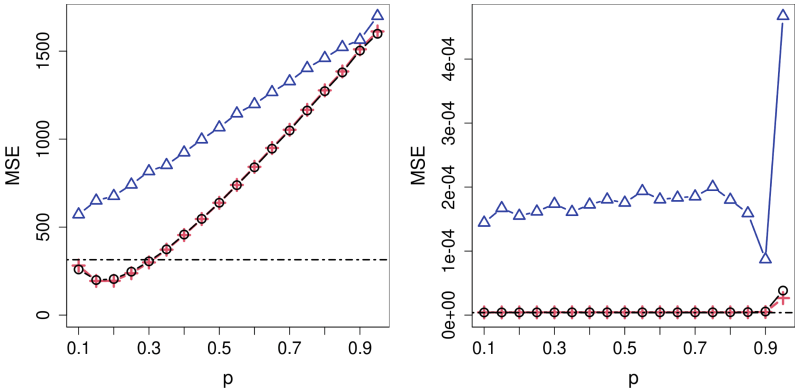


Fig. 1. Parameter estimation results based on 500 realisations of a hard-core model on $S = [0,1]^2$ with parameters $R_0 = 0.05$ and $\beta_0 = 100$. Left: MSE($\widehat{\beta}$). Right: MSE($\widehat{R}$). Loss functions: $\mathcal{L}_1$ (black line with circles), $\mathcal{L}_2$ (red line with plus signs) and $\mathcal{L}_3$ (blue line with triangles), using the test function $h_\theta(u;\mathbf{x}) = 1/(p\lambda_\theta(u;\mathbf{x}))$, in combination with MCCV, where $p = 0.1, 0.15, \ldots, 0.95$ and $k = 100$. The dotted black lines represent pseudolikelihood estimation.

best performance in terms of MSE($\widehat{\beta}$). More specifically, when $p = 0.1$ the value of MSE($\widehat{\beta}$) for $\mathcal{L}_1$ is given by 199.4225 and for $\mathcal{L}_2$ it is given by 194.2291, which are significantly lower than what pseudolikelihood gives rise to. Also here, $\mathcal{L}_3$ performs much worse than its three competitors, as a consequence of a higher bias. This inferior performance may, however, be model specific.

5 Conclusions

We have illustrated that Point Process Learning outperforms pseudolikelihood estimation in the context of fitting the hard-core Gibbs model. Based on the findings in our simulation study, the loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$ in combination with MCCV and $p < 0.3$, which perform next to identically, are the preferred choices here. Even more interestingly, perhaps, is that Point Process Learning does not seem to require a plug-in/profile approach when there are identifiability issues present (a fixed estimate of R_0 does not need to be plugged into (4) to obtain a good estimate of β_0).

References

1. Baddeley, A., Rubak, E., Turner, R.: Spatial Point Patterns: Methodology and Applications with R. CRC Press, Boca Raton (2015)
2. Chiu, S.N., Stoyan, D., Kendall, W.S., Mecke, J.: Stochastic Geometry and its Applications. Wiley, Hoboken (2013)

3. Cronie, O., Moradi, M., Biscio, C.A.: A cross-validation-based statistical theory for point processes. Biometrika **111**(2), 625–641 (2024)
4. van Lieshout, M.N.M.: Markov Point Processes and Their Applications. Imperial College Press/World Scientific (2000)

Temporal Nearest Neighbor Gaussian Process (tNNGP) with Flexible Covariance for Modelling Physical Activity

Marco Mingione[1(✉)] and Pierfrancesco Alaimo Di Loro[2]

[1] Department of Political Sciences, Roma Tre University, 00154 Rome, Italy
marco.mingione@uniroma3.it
[2] Department of GEPLI, Libera Università Maria Ss. Assunta (LUMSA), 00192 Rome, Italy
p.alaimodiloro@lumsa.it

Abstract. We propose a generalization of the Temporal Nearest Neighbor Gaussian Process to model high-frequency data in the context of individual physical activity monitoring. To our knowledge, previous applications considered the exponential as a model for the temporal covariance but more flexible alternatives are readily available and could be easily embedded in the same modelling framework. Thus, the idea is to verify the comparative performances of alternative temporal covariance specifications and find the one that best suits applications in such a context. The comparison is pursued on a dataset compiled from the PASTA-LA study, which monitored the physical activity patterns of individuals across space and time using a tri-axial accelerometer.

Keywords: NNGP · covariance · sparsity · Gaussian processes

1 Introduction

The monitoring of human activity patterns has generated substantial interest among experts in public health. Indeed, understanding the dynamics and factors influencing Physical Activity (PA) patterns could help gear interventions toward social and environmental features related to higher (or lower) PA. Luckily, non-invasive technologies for monitoring spatial energetics and promoting physical activity continue to emerge, such as wearable devices equipped with ad-hoc sensors that can record repeated measurements of time, location and PA endpoints at a very high resolution [6]. Among these, accelerometers (increasingly conspicuous because of their affordability and availability in common-use devices) measure the acceleration along different axes and can be elaborated to provide proxies of body movement (hence PA, [1,4]).

Analyzing these kinds of data and accounting for their full spatio-temporal structure can be particularly challenging because of their (typically) massive

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 212–217, 2025.
https://doi.org/10.1007/978-3-031-64346-0_36

size. On top of that, they also present substantial unobserved heterogeneity and intrinsic temporal dependence [2]. While these last two aspects are often neglected, [1] tackles these issues by devising a model for individual-wise trajectories of PA and considering a combination of spline smoothing (for the spatial effects) and Gaussian Processes (for the temporal dependence). The *Big-Data* problem is then handled by exploiting the low-rank and the NNGP approximations, as reported in Sect. 2.

While general in the model specification, one main limitation of the original work lies in the consideration of the exponential function as the only viable temporal covariance model in the proposed application. However, there is no guarantee that such a choice will provide the best results in terms of model fitting and prediction performances. We here extend the original application by considering alternative temporal covariance models (Sect. 2.1), whose merits are compared in terms of goodness-of-fit penalized criterion and computational performances (Sect. 3). The comparison is pursued on a dataset compiled from the same PASTA-LA study, which monitored the physical activity patterns of a group of individuals using a GPS and a tri-axial accelerometer.

2 Method

Let $\mathcal{T} = \{t_i\}_{i=1}^n$, with $t_i \in \mathbb{R}^+$ be the set of the n observed time points. We model $\boldsymbol{Y}(\mathcal{T})$ as the finite realization of a uni-variate process $\boldsymbol{Y}(\cdot)$ over $\mathbb{R}^+$ as

$$Y(t) = \boldsymbol{X}(t)^\top \boldsymbol{\beta} + w(t) + \varepsilon(t), \quad t \in \mathbb{R}^+, \tag{1}$$

where $\boldsymbol{X}(t)$ is a $n \times p$ covariate matrix, $\varepsilon(t) \sim \mathcal{N}(0, \tau^2)$, $\tau^2 \in \mathbb{R}^+$, is a white noise process for measurement error, and $w(t) \sim \mathcal{GP}(0, c_{\boldsymbol{\theta}}(\cdot,\cdot))$, where $c_{\boldsymbol{\theta}}(\cdot,\cdot)$ is a covariance function with parameters $\boldsymbol{\theta} \in \Theta$.

Equation (1) leads to a hierarchical model with posterior distribution

$$p(\boldsymbol{\beta}, w, \boldsymbol{\theta}, \tau^2 \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\theta}, \tau^2) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(w \,|\, \boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + w, \tau^2 \boldsymbol{I}). \tag{2}$$

Applying (2) involves the determinant and inverse of $\boldsymbol{C}_{\boldsymbol{\theta}}$, which require $\mathcal{O}(n^2)$ storage space and $\mathcal{O}(n^3)$ floating point operations (flops). For these reasons, the estimation process is fast and feasible as long as the data size is relatively small, i.e. smaller than 10^3. When this is not the case, the computational burden increases, limiting the straightforward application of such modelling tools. To solve this issue, we adapt [7] approximation to the random effect $w(\cdot)$. Beginning with the observed time points $\{t_1 < t_2 < \cdots < t_n\}$ and the directed acyclic graphical (DAG) representation $p(w) = p(w_1) \prod_{i=2}^n p(w_i \,|\, w_1, \ldots, w_{(i-1)})$, we define

$$p(w) \approx \widetilde{p}(w) = p(w_1) \prod_{i=2}^{n} p(w_i | w_{N(i)}) \,, \tag{3}$$

where $\widetilde{p}(\cdot)$ is the joint density derived from $p(w)$ by restricting the parents (conditional sets) of each w_i in the DAG to a set $w_{N(i)} = \{w_j : j \in N(i)\}$, where $N(i)$ is a set of prefixed size m comprising the m nearest neighbors of t_i from the

past. Thus, $N(i) = \{t_{(i-m)}, \dots, < t_{(i-1)}\}$ for $i > m$ and $N(i) = \{t_1, \dots, t_{(i-1)}\}$ for $i \leq m$. Such approximations yield valid probability likelihoods and can be extended to stochastic processes for inference on arbitrary time points [3,5].

The connection between sparsity and conditional independence follows by writing (3) as a linear model $w = \boldsymbol{A}w + \boldsymbol{\eta}$, where $\boldsymbol{A}$ is a $n \times n$ strictly lower triangular matrix, $\boldsymbol{\eta} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{D}$ is the $n \times n$ diagonal matrix such that $[\boldsymbol{D}]_{ii} = d_{ii} = \text{Var}\left(w_i | \{w_j, j < i\}\right)$ for $i = 1, \dots, n$ [3]. The DAG imposes the lower-triangular structure on $\boldsymbol{A}$ and its (i, j)-th entry is allowed to be nonzero only for $j \in N(i)$. Therefore, each row of $\boldsymbol{A}$ has at most m nonzero entries so that $\widetilde{\boldsymbol{C}}^{-1} = (\boldsymbol{I} - \boldsymbol{A})^\top \boldsymbol{D}^{-1} (\boldsymbol{I} - \boldsymbol{A})$ is sparse, where $\widetilde{\boldsymbol{C}}^{-1}$ is the precision matrix corresponding to $\widetilde{p}(w)$. Replacing $\boldsymbol{C}$ with $\widetilde{\boldsymbol{C}}$ in (2) yields a computationally efficient hierarchical model with a prior $w \sim NNGP(\boldsymbol{0}, \widetilde{\boldsymbol{C}})$.

The key observation is that the nonzero elements of the i-th row of $\boldsymbol{A}$ are the solution $\boldsymbol{a}$ of the $m \times m$ linear system $\boldsymbol{C_\theta}[N(i), N(i)]\boldsymbol{a} = \boldsymbol{C_\theta}[N(i), i]$, where $[\cdot, \cdot]$ indicates submatrices defined by the given row and column index sets. Obtaining the nonzero elements of $\boldsymbol{A}$ and $\boldsymbol{D}$ is cheaper than it would have been without sparsity. This also delivers the quadratic form $w^\top \widetilde{\boldsymbol{C}}^{-1} w$ in terms of $\boldsymbol{A}$ and $\boldsymbol{D}$ and the determinant $\det(\widetilde{\boldsymbol{C}}) = \prod_{i=1}^n d_{ii}$ at almost no additional cost. The lower triangular matrix $\boldsymbol{A}$ is not just sparse but also banded, with a lower bandwidth equal to m. Consequently, $\widetilde{\boldsymbol{C}}^{-1}$ is also banded with lower and upper bandwidth equal to m. This leads to further accrual of computational benefits. The overall cost is $\mathcal{O}(nm^3)$ (linear in n) for computing the posterior for any given values of the parameters.

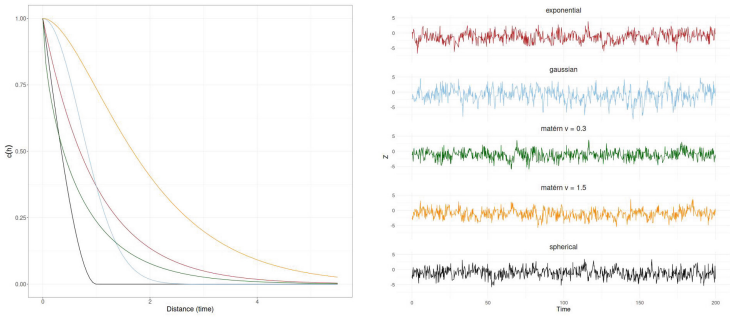
2.1 Covariance Functions

Without loss of generality, denoting with $h = |t_i - t_{i'}| > 0$ the distance in time between subsequent observations, we can express any covariance function as $c(h) = \sigma^2 \cdot \rho(h)$, where $\rho(\cdot)$ is a correlation function such that $\rho(0) = 1$ and $\rho(h) \overset{h \to \infty}{\to} 0$, and where σ^2 represents the variance of the process.

We here introduce 4 of the most widely used models in terms of the corresponding correlation function $\rho_\theta(\cdot)$, with decay parameter $\phi > 0$.

- **Exponential:** $\rho_\theta(h) = \exp\{-\phi \cdot h\}$
- **Gaussian:** $\rho_\theta(h) = \exp\left\{-\phi^2 \cdot h^2\right\}$
- **Matérn:**
$$\rho_\theta(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\sqrt{2\nu} \cdot \frac{h}{\phi}\right) \cdot \mathcal{K}_\nu \left(\sqrt{2\nu} \cdot \frac{h}{\phi}\right),$$
where $\Gamma(\cdot)$ is the gamma function, ν is a positive smoothness parameter and $\mathcal{K}_\nu$ is the modified Bessel function of second type. Exponential and Gaussian are particular cases of the Matérn for $\nu = 1/2$ and $\nu \to \infty$, respectively.
- **Spherical:**
$$\rho_\theta(h) = \begin{cases} 1 - \frac{3 \cdot h}{2 \cdot \phi} + \frac{h^3}{2 \cdot \phi^3} & \text{if } 0 < h \leq \phi \\ 0 & \text{if } h > \phi \end{cases}$$

As an example, we simulate $n = 1000$ observations of a process with 0 mean and for five different specifications of the covariance, where $\sigma^2 = \phi = 1$ in all cases and $\nu = 0.3, 1.5$ for the Matérn cases. The theoretical covariance functions are reported in Fig. 1a and describe various scenarios with the correlation decreasing to 0 at a distance between subsequent observations ranging from 1 (spherical) to more than 5 (Matérn with $\nu = 1.5$) time units. The corresponding data are shown in Fig. 1b.



(a) (b)

Fig. 1. (a) Example of theoretical covariance functions; (b) simulated data, corresponding to each one of the considered covariance specifications: exponential (red), gaussian (blue), matérn with $\nu = 0.3$ (green), matérn with $\nu = 1.5$ (yellow) and spherical (black).

Table 1. DIC values and parameters' estimates.

	Exponential	Matérn	Gaussian	Spherical
DIC	2081.60	3769.16	4745.05	1268.32
ϕ	0.82 (0.26, 1.85)	0.50 (0.10, 1.19)	3.96 (2.95, 4.92)	0.62 (0.25, 1.19)
σ^2	0.79 (0.27, 1.74)	1.38 (0.42, 4.05)	1.70 (0.86, 2.90)	0.63 (0.27, 1.32)
τ^2	0.008 (0.007, 0.009)	0.008 (0.007, 0.009)	0.010 (0.009, 0.011)	0.008 (0.007, 0.009)
ν	–	0.50 (0.42, 0.60)	–	–

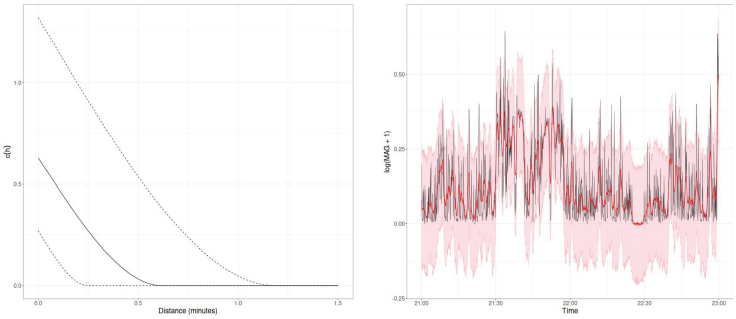
3 Application

For estimation purposes, we use data from just one individual collected from 9 p.m. to 11 p.m., for a total of $n = 1438$ observations. The individual is a 27-year-old female Asian with a body mass index of 24 (normal weight).

We estimate the model described in Sect. 2 for each specification of the temporal covariance defined in Sect. 2.1. The prior setting for the parameters is the following:

$$\sigma^2 \sim IG(2,2) \qquad \tau^2 \sim IG(2,2) \qquad \phi \sim Unif(0.5,5) \qquad \nu \sim Unif(0.01,2)$$

For each model, we run one chain of 20,000 iterations, half burn-in and thinning equal to 10. As suggested in the seminal paper by [3], we set the number of neighbours equal to 15, which has been proven to be sufficient in well approximating the *true* Gaussian process for $w(t)$. Results are reported in Table 1 where we show the DIC of each model and the posterior parameters' estimates. The covariance's specification yielding the best fitting performances is the spherical one, with a range equal to $\hat{\phi} = 0.62\,(IC_{0.95} = [0.25, 1.19])$, namely the correlation between subsequent observations decays to zero after approximately half a minute. The estimated covariance function is shown in Fig. 2a, together with the 95% credible intervals. The fitting performances of the model with the spherical covariance are instead illustrated in Fig. 2b, where we notice that the average movement pattern is almost perfectly recovered.



(a) (b)

Fig. 2. (a) Estimated spherical covariance (black solid line) with the 95% credible intervals (black dashed lines); (b) observed (black solid line) and estimated values (red solid line) with the 95% prediction intervals (red area).

4 Conclusions

The study of correlation patterns driving the dynamics of PA in humans could enhance our understanding of individual behaviour and contribute to more effective personalized interventions aimed at improving public health outcomes. Here, we have integrated different covariance specifications into the modelling framework proposed by [1] to model the PA of an individual during daytime in a free-living environment, and showed that results could differ according to the

chosen covariance model. Further analysis of such data should account for individual time-varying covariates that could possibly affect the PA levels during the day (e.g. distance from the workplace, closeness to green areas, etc.).

Acknowledgements. This work has been supported by MIUR, grant number 2022XRHT8R - The SMILE project: Statistical Modelling and Inference for Living the Environment.

References

1. Alaimo Di Loro, P., Mingione, M., Lipsitt, J., Batteate, C.M., Jerrett, M., Banerjee, S.: Bayesian hierarchical modeling and analysis for actigraph data from wearable devices. Ann. Appl. Stat. **17**(4), 2865–2886 (2023)
2. Bai, J., Sun, Y., Schrack, J.A., Crainiceanu, C.M., Wang, M.-C.: A two-stage model for wearable device data. Biometrics **74**(2), 744–752 (2018)
3. Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E.: Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. **111**(514), 800–812 (2016)
4. Doherty, A., et al.: Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. PloS One **12**(2) (2017)
5. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
6. Sikka, R.S., Baer, M., Raja, A., Stuart, M., Tompkins, M.: Analytics in sports medicine: implications and responsibilities that accompany the era of big data. JBJS **101**(3), 276–283 (2019)
7. Vecchia, A.V.: Estimation and model identification for continuous spatial processes. J. Roy. Stat. Soc.: Ser. B (Methodol.) **50**(2), 297–312 (1988)

Spatio-temporal Clustering of $PM_{2.5}$ in northern Italy using a Bayesian model

Florian Wolf[1(✉)], Alessandro Carminati[2], and Alessandra Guglielmi[2]

[1] Department of Mathematics, Technical University Darmstadt, Darmstadt, Germany
florian.wolf@stud.tu-darmstadt.de

[2] Department of Mathematics, Politecnico di Milano, Milan, Italy
{alessandro.carminati,alessandra.guglielmi}@polimi.it

Abstract. Northern Italy is a well-known hotspot of air pollution in Europe, due to its high population density, characteristic geography and specific climate conditions. Consequently, there is a considerable interest in investigating the temporal patterns of air quality across numerous stations scattered throughout the region. In this work, we apply a Bayesian spatio-temporal product partition model for clustering particulate matter ($PM_{2.5}$) concentration in northern Italy and compare its performance with a baseline spatial-only model.

Keywords: Bayesian Inference · Clustering · Product Partition Models

1 Introduction

Air pollution is the contamination of air due to emission of gases and particulate matter (PM) into the atmosphere. Due to their significant risk to human health and the ecosystem, it is of great interest studying fine dusts, i.e. particulate matters with a diameter of $2.5\,\mu m$ or smaller, namely $PM_{2.5}$, mostly caused by vehicular emissions, domestic heating, agricultural activities etc. Understanding the temporal patterns of $PM_{2.5}$ levels is crucial for identifying trends, potential sources and developing effective pollution control strategies. For more detailed information on $PM_{2.5}$ and air pollution in Europe in general, we refer to [2].

Northern Italy is a well-known and intensively studied hotspot of PM pollution in Europe caused by a combination of (a) a high population density, high level of urbanization and industrial processes and (b) the characteristic shape and climate conditions that prevent effective dispersion. Inherently, PM pollution data are *spatial* due their intrinsic geographical characterization and usually are *time-dependent*, since they are measured at different times.

In this paper, we focus on data which are spatially correlated $PM_{2.5}$ measurements time series in monitoring stations. Interest is in clustering the monitoring stations to highlight similar sites in terms of $PM_{2.5}$ pollution, aiming to enhance our understanding of air pollution dynamics and discern geographical patterns

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 218–223, 2025.
https://doi.org/10.1007/978-3-031-64346-0_37

of pollution concentration. We apply the spatiotemporal model developed by Page et al. [8] to cluster $PM_{2.5}$ pollution data.

2 The Agrimonia Dataset and Data Preprocessing

The dataset we analyze is part of the Agrimonia dataset (see [10]) which integrates satellite data, model output and in-situ measurements sourced from national and international agencies, each with varying spatial and temporal resolutions. After aggregating $PM_{2.5}$ to have weekly measurements, in the log scale, we obtain of $n = 46$ monitoring stations (see Fig. 2 or 3 for their locations) for $T = 52$ weeks of year 2019. Temporal autocorrelation analysis highlights strong temporal dependence of the $\log(PM_{2.5})$ concentration, in particular the average lag 1 autocorrelation is 0.54 ± 0.04 (95% confidence interval).

3 The Bayesian Model

For all the weeks in 2019, we introduce the sequence of parameters able to give time-dependent cluster estimates of the n monitoring stations based on their $\log(PM_{2.5})$ concentration values. We denote by $\rho_t = \{S_{1t}, \ldots, S_{k_t,t}\}$ the partition of the stations into k_t clusters at week t, for $t = 1, \ldots, T$. We assume the spatio-temporal Random Partition Model (stRPM) prior for $(\rho_1, \ldots, \rho_T)$, which was recently introduced in [8]. This prior models the sequence of partitions with a first-order Markovian structure, i.e., $\pi(\rho_t \mid \rho_{t-1}, \ldots, \rho_1) = \pi(\rho_t \mid \rho_{t-1})$ for $t = 2, \ldots, T$.

Let $i = 1, \ldots, n$ be the index of the monitoring stations that record $\log(PM_{2.5})$. For $t = 1, \ldots, T$, we introduce the auxiliary variable $\boldsymbol{\gamma}_t = (\gamma_{1t}, \ldots, \gamma_{nt})$, which guides the similarity between ρ_t and ρ_{t-1}:

$$\gamma_{it} = \begin{cases} 1, \text{ station } i \text{ is } \textbf{not} \text{ reallocated from time } t-1 \text{ to } t \\ 0, \text{ else} \end{cases} \tag{1}$$

Then the prior assumes that $\gamma_{it} \mid \alpha_t \overset{ind.}{\sim} \text{Be}(\alpha_t)$ for each $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\text{Be}(\cdot)$ denotes the Bernoulli distribution. The temporal dependence parameters $(\alpha_1, \ldots, \alpha_T)$ are assumed independent and identically distributed from the $\text{Beta}(a_\alpha, b_\alpha)$ distribution. The marginal prior of the random partition of the first week, $\pi(\rho_1)$, is a spatial Product Partition Model (sPPM) [6], which is proportional to $\prod_{j=1}^{k_t} c(S_{jt}) g(\boldsymbol{s}^\star_{jt})$, where $c(\cdot)$ is the cohesion function, which produces cluster weights, and $g(\cdot)$ is the similarity function, which measures the compactness of the spatial coordinates in $\boldsymbol{s}^\star_{jt}$. It can be proven (see Proposition 1 and Sect. 4.2 of [8]) that marginally $\rho_1, \ldots, \rho_T$ are identically distributed with law coming from the sPPM used to model ρ_1. Therefore, for each week, the similarity function introduces spatial information about the stations in the clustering process.

We consider the cohesion function $c(S_{jt}) = M \times (|S_{jt}| - 1)!$, where M is an hyperparameter called concentration or total mass. This function is connected with the Dirichlet process [3]. We model $g(\cdot)$ with a double dipper similarity function [9].

Our model can be described as follows. For each $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let Y_{it} be the $\log(\text{PM}_{2.5})$ concentration value in station i at week t, and let c_{it} be its clustering label: $c_{it} = j$ means that station i in week t belongs to cluster S_{jt}. Based on the temporal autocorrelation analysis presented in Sect. 2, we model the observations from a single station with an AR(1) structure:

$$\begin{aligned} Y_{it} \mid \boldsymbol{\mu}_t^\star, \sigma_t^{2\star}, \boldsymbol{c}_t &\overset{ind.}{\sim} \mathcal{N}(\mu^\star_{c_{it}t} + \eta_i Y_{it-1}, \sigma^{2\star}_{c_{it}t}) \quad i = 1, \ldots, n,\ t = 2, \ldots, T \\ Y_{i1} \mid \boldsymbol{\mu}_1^\star, \sigma_1^{2\star}, \boldsymbol{c}_1 &\overset{ind.}{\sim} \mathcal{N}(\mu^\star_{c_{i1}1}, \sigma^{2\star}_{c_{i1}1}) \quad i = 1, \ldots, n \end{aligned} \tag{2}$$

where η_i is a unit specific temporal dependence parameter, and $\boldsymbol{\mu}_t^\star$ and $\boldsymbol{\sigma}_t^{2\star}$ are vectors containing cluster-specific mean and variance values, respectively. We assume a Laplace prior for η_i and a hierarchical model prior for $\boldsymbol{\mu}_t^\star$ and $\boldsymbol{\sigma}_t^{2\star}$:

$$\begin{aligned} \text{Logit}(0.5(\eta_i + 1)) &\overset{i.i.d.}{\sim} \text{Laplace}(a, b) \quad i = 1, \ldots, n \\ (\mu^\star_{jt}, \sigma^\star_{jt}) \mid \theta_t, \tau_t^2 &\overset{ind.}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \quad t = 1, \ldots, T,\ j = 1, \ldots, k_t \\ (\theta_t, \tau_t) &\overset{i.i.d.}{\sim} \mathcal{N}(\phi_0, \lambda^2) \times \mathcal{U}(0, A_\tau) \quad t = 1, \ldots, T \\ (\phi_0, \lambda) &\overset{i.i.d.}{\sim} \mathcal{N}(m_0, s_0^2) \times \mathcal{U}(0, A_\lambda) \end{aligned} \tag{3}$$

where $a, b, A_\sigma, A_\tau, A_\lambda, m_0, s_0^2$ are hyperparameters. We assume the clustering labels are obtained from the stRPM prior, described at the beginning of this section:

$$\begin{aligned} c_{it} &\overset{i.i.d.}{\sim} \text{stRPM}(\boldsymbol{\alpha}, M) \quad i = 1, \ldots, n;\ t = 1, \ldots, T \\ \alpha_t &\overset{i.i.d.}{\sim} \text{Beta}(a_\alpha, b_\alpha) \quad t = 1, \ldots, T \end{aligned} \tag{4}$$

Summing up, the model we assume for our data is (2), (3) and (4).

4 Posterior Inference

We apply (2), (3) and (4) to the log of weekly averages of $\text{PM}_{2.5}$ in monitoring stations in northern Italy in 2019. We use 22,000 MCMC iterations, a burn-in of 2,000 iterations and a thinning of 10, resulting in 2,000 samples. Hyperparameters are fixed as follows: $m_0 = 2.91$ and $s_0^2 = 200$, based on the mean and variance of the logarithm of previous year's (2018) data, and $A_\sigma = 0.1, A_\tau = 1.0, A_\lambda = 1.0, a = 0, b = 1.0, a_\alpha = 1.0, b_\alpha = 1.0, M = 0.1$, determined from the experiments presented in [8]. The cluster estimates of the stations' labels are obtained by minimizing the posterior expectation of the Binder loss function using the `salso` package [1].

As a benchmark, we modify the stRPM by fixing $\alpha_t = 0$ for each $t = 1, \ldots, T$, obtaining a model where the partitions in different weeks are considered a priori independent. We keep the same hyperparameters' values and MCMC settings. With a slight abuse of notation, in the following we call this model sPPM.

Comparison to the sPPM-Benchmark. Table 1 presents three predictive goodness-of-fit indices for both models: log pseudo marginal likelihood (LPML), widely applicable information criterion (WAIC) [5], and mean squared error (MSE). The latter was computed considering the estimates of the posterior means as predictors.

The indices show that relaxing the time-independency assumption enhances the predictive efficacy of the model, albeit at the expense of increased computational time by an order of magnitude. Notably, the MSE is approximately tenfold higher for the time-dependent prior, a discrepancy attributable to the model's utilization of a greater number of clusters.

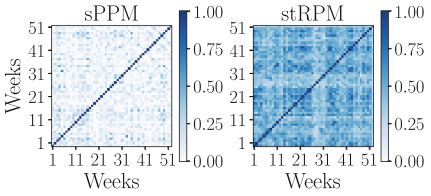


Fig. 1. Lagged Adjusted Rand Index (ARI) for the partition estimate.

Table 1. Predictive goodness-of-fit metrics (weekly average) and computational time.

	sPPM	stRPM
LPML	17.55	**27.12**
WAIC	29.19	**14.79**
MSE	**0.12**	1.68
Time [s]	**20.35**	287.24

Our analysis reveals that the sPPM favors fewer yet larger clusters, while the stRPM tends to utilize a slightly larger number of clusters throughout the year. The lagged Adjusted Rand Index (ARI) values in Fig. 1 show that temporal dependence in the stRPM prior allows for a smoother evolution of the cluster estimates over time.

Figure 2 and 3 show the cluster estimates of the monitoring stations for two consecutive weeks of the year 2019, obtained using stRPM. Despite some cluster reallocations, one observes the smoothing behavior induced by the temporal information as the cluster estimates for most of the stations do not change. Furthermore, we see that the two stations in Milan are intuitively correctly clustered together with other cities of high pollution like Brescia and Bergamo.

Posterior Covariate Analysis. Since the stRPM model is covariate-agnostic, we analyze the empirical distribution of some covariates within the estimated clusters at weeks 6 and 7. Specifically, we focus on the maximum Planetary Boundary Layer Height (BLH_{max}) and Total Precipitation, which are known to strongly influence particulate matter (PM) concentration. Figure 4 displays the distribution of these two covariates for the three main clusters (excluding singletons)

Fig. 2. Clustering of the stRPM model for **week 6/52** with 6 clusters in total.

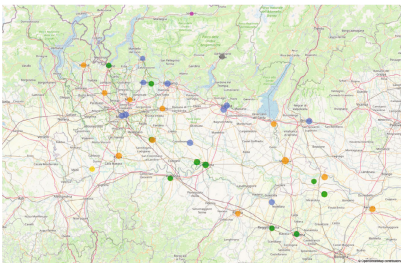


Fig. 3. Clustering of the stRPM model for **week 7/52** with 6 clusters in total.

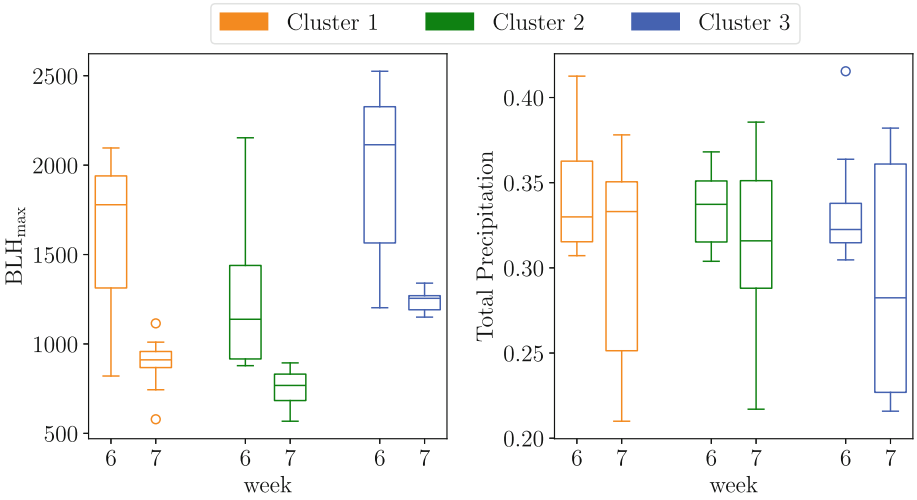


Fig. 4. Covariate distribution of the maximum Planetary Boundary Layer Height (BLH_{max}) and the Total Precipitation for the **weeks 6 and 7**. The color scheme is coherent with Fig. 2 and 3. Singleton clusters are omitted.

from Figs. 2 and 3, using a coherent color scheme. In both weeks, the distributions of BLH_{max} values in the cluster estimates are clearly different, while the distributions of Total Precipitation values are similar. Our analysis reveals that BLH_{max} values exhibit the highest correlation with log-particulate matter concentration. This result confirms the model's implicit dependency and aligns with our preliminary data analysis. However, we anticipate that including covariate information (e.g., Total Precipitation) in the partition model will improve the clustering result and further enhance cluster separation.

Future Work. As discussed in Sect. 1, $PM_{2.5}$ concentration is heavily connected to the altitude of the stations, weather phenomena and the level of urbanization of the surroundings causing the emissions [4]. So far, covariate-informed product

partition models (PPMx) priors have been proposed in the literature (see [7]), but these models rely on the strong assumption of spatial and/or temporal independence of the measurements. Thus, future work includes proposing priors for clustering spatial locations over time, but also including extra covariates.

Acknowledgments. We would like to additionally thank Elisa Borrini, Filippo Carbonara, Benedetta Cefaloni, Dina Sophie Ettel and Alessandro Grignani, M.Sc. students in Mathematical Engineering at Politecnico di Milano, for the implementation of the code. Alessandro Carminati and Alessandra Guglielmi have been partially supported by MUR, Grant Dipartimento di Eccellenza 2023–2027.

References

1. Dahl, D.B., Johnson, D.J., Müller, P.: Salso: search algorithms and loss functions for Bayesian Clustering (2022). R package version 0.3.35
2. European Environment Agency: Harm to human health from air pollution in Europe: burden of disease 2023. EEA Briefing. Publications Office (2023)
3. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973). https://doi.org/10.1214/aos/1176342360
4. Frigeri, M., Guglielmi, A., Lonati, G.: A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy. In: Book of Short Papers - Statistical Learning, Sustainability and Impact Evaluation, pp. 1009–1014. Pearson (2023)
5. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat. Comput. **24**(6), 997–1016 (2013). https://doi.org/10.1007/s11222-013-9416-2
6. Page, G.L., Quintana, F.A.: Spatial product partition models. Bayesian Anal. **11**(1), 265–298 (2016). https://doi.org/10.1214/15-BA971
7. Page, G.L., Quintana, F.A.: Calibrating covariate informed product partition models. Stat. Comput. **28**, 1009–1031 (2017)
8. Page, G.L., Quintana, F.A., Dahl, D.B.: Dependent modeling of temporal sequences of random partitions. J. Comput. Graph. Stat. **31**, 614–627 (2022). https://doi.org/10.1080/10618600.2021.1987255
9. Quintana, F.A., Müller, P., Papoila, A.L.: Cluster-specific variable selection for product partition models. Scand. J. Stat. **42**(4), 1065–1077 (2015)
10. Rodeschini, A.F.J., et al.: Agrimonia: a dataset on livestock,meteorology and air quality in the Lombardy region, Italy. Sci. Data **10**(1) (2023). https://doi.org/10.1038/s41597-023-02034-0

Graphs as Unifying Logical Structures in the Construction of Information System for Complex Domains

Cristina Martelli(✉), Adham Kahlawi, and Maria Flora Salvatori

Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, 50134 Florence, Italy
{cristina.martelli,adham.kahlawi,mariaflora.salvatori}@unifi.it

Abstract. In an era marked by widespread data generation and the accessibility of advanced technologies, one may wonder about the relevance of standard coding. Is it still crucial to invest in the development and application of standardized coding frameworks? This paper posits that the key to addressing these inquiries lies in acknowledging and augmenting the inherent logical structure of the graph, which is omnipresent and integral at all stages in the lifecycle of a robust, detailed, high-quality, and statistically reusable data source.

Keywords: Metadata Information Systems · Knowledge Graphs

1 Introduction

The construction of a data source [1, 2] typically entails several critical steps: (i) analyzing and collecting the domain of interest's description, (ii) producing the conceptual structure of the source, (iii) translating the conceptual model into a technological tool, (iv) utilizing the tool to intervene in the problem domain (management phase), and (v) understanding and making decisions based on it (informative-statistical phase). These steps must be undertaken with the objectives of source robustness and evolution in mind. To achieve the previously mentioned goals, it is imperative that these diverse phases are deeply interconnected and harmonized. This ensures that (i) domain experts and problem specialists feel heard and see their narratives reflected in the technical system, (ii) managers and researchers access data with a level of detail and interconnectedness that yields meaningful information, and (iii) the system is capable of expansion as the understanding of reality becomes more comprehensive. Unfortunately, this ideal operational environment is often unattainable, partly due to the segmentation of these construction steps among various professions. These professions, often working in silos, may lack the ability for mutual, active listening.

Central to addressing these challenges is a common logical structure: the graph. In this work, we aim to demonstrate how the logical structure of the graph, when recognized and consciously managed, facilitates a smooth transition between the different construction phases of data sources.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 224–229, 2025.
https://doi.org/10.1007/978-3-031-64346-0_38

The structure of the work is as follows: discussion on the role of graphs across all stages of a data source's lifecycle (Sects. 2.1, 2.2), evolution of a data source within the context of information systems that rely on metadata networks (Sect. 2.3), with particular reference to the role of semantic ontologies in the construction of metadata information systems (Sect. 2.1.5). Finally, (Sect. 3), an illustrative example of a metadata-based information system is presented and analyzed to demonstrate its utility in facilitating the creation of graphs essential for all stages of source development.

2 The Graph in the Context of Data Steps Biography

Within the scope of this work, we adhere to the definition of a graph as a first-order logical model [3–5]. In this model, nodes represent entities (objects) within the discourse domain, arcs symbolize the relationships between these entities, and the properties of both entities and relationships can be elucidated through first-order logical formulas. This broader definition, as we will demonstrate, is not only well-suited to the initial phases of source generation—specifically, the manipulation of domain experts' narratives—but also guides us towards a more formalized interpretation.

The significance of employing the concept of a graph within first-order logical models is attributable to the role of first-order formal languages in semantics. These languages provide a means to ascertain the truth or falsity of a formula within a given context. This capability finds application in various fields, including automatic software verification and database design. The latter, in particular, will be revisited and examined in detail when we discuss the physical generation of administrative archives that gather data on a specific domain.

Domain experts often articulate their understanding of the in a discursive and unstructured manner. This specific mode of expression leads them, and their technical counterparts, to view domain experts as somewhat external to the process of generating and computerizing a source. Commonly, their input is limited to an initial description and they are not involved in the subsequent steps until the system is ready for use.

However, when their narrative is transformed into structured sentences—comprising a subject, verb (or predicate), and object—the original, unstructured linguistic domain description is converted into a graph structure:

a. Subject: The subject of the sentence represents an entity within the discourse domain. In logical terms, the subject can be interpreted as a variable that assumes values within the domain.
b. Verb/Predicate: The verb in the sentence denotes a relationship between the subject and another entity, namely the object. From a logical perspective, the verb functions as a predicate that connects the subject to another variable or constant.

This process not only facilitates a formal approach to the description but also underscores the importance of involving domain experts throughout the development cycle, enabling a more accurate and comprehensive representation of the domain in question.

2.1 Integrating Graphs: Sources for the Study of Complexity

The first step for the construction of a source, right from its most narrative and discursive dimension, can be represented with a graph-type structure (like the classical Entity/Relationship model), which allows the domain expert to control the structure of the system that will be built and to check, together with the technicians, the performance of the system from the point of view of queries and logical inference. Point of view of queries and logical inference. Furthermore, the representation in terms of graph and first-order logic allows technicians to rely on a formalized language that can be directly translated into storage and semantic structures.

The management of domain experts' narratives in the form of a graph supports also the evolution of sources, i.e., the conditions for integrating graphs representing a specific domain of discourse to outline a broader information source. This aspect typifies the reuse of administrative sources when attempting to broaden the information spectrum by integrating (often with difficulty) different databases.

Among the conditions for Integrating Graphs into a Larger Graph Structure, we will particularly focus on the Semantic Alignment [6], intended as the (i) correspondence between concepts and relationships represented in different graphs; (ii) ontologies or controlled vocabularies to unify terminology and, finally (iii) semantic mappings to make clear the relationships between homologous concepts.

To achieve semantic alignment, standard encodings such as NACE or ISCO, are traditionally employed to characterize the nodes, thereby facilitating the reciprocal integration of graphs. However, this work proposes an expansion beyond the conventional hierarchical structure of these standards. Instead, it advocates for a focus on metadata within information systems [7], suggesting a more flexible and comprehensive approach to encoding and integration. This adjustment aims to better capture the nuanced relationships and concepts within and across various domains, enhancing the utility and accuracy of the constructed information sources.

2.2 MIS and Ontologies

Significant advancements in the approach to Management Information Systems (MIS) design and construction, have been achieved through the methodology and technology of ontologies [8]. An ontology, defined as a formal representation (utilizing first-order logic) of a domain's conceptualization, offers a vocabulary of terms and interrelationships, thereby facilitating the description and reasoning within a specific area of knowledge. This capability to formalize and articulate knowledge domains makes ontologies instrumental in enhancing data sharing, interoperability, knowledge discovery, reasoning, decision-making, and ultimately, improving efficiency and productivity.

Ontologies find application across a broad spectrum of fields, including the semantic web, data integration, knowledge management, natural language processing, and bioinformatics. By providing a structured vocabulary for metadata, ontologies play a pivotal role, especially when adopting traditional standard coding. This feature is particularly beneficial for achieving a finer granularity of understanding and interpretation within and across various domains.

Typically, ontologies are published on the web, thereby facilitating the dissemination and adoption of a shared conceptualization. This openness supports collaborative efforts and contributes to the establishment of common understandings across different entities and disciplines.

Ontologies are underpinned by a graph structure, (subject-verb/predicate-object). The reliance on this graph structure not only aids in maintaining coherence in conceptualization but also in the integration and evolution of knowledge bases, making ontologies a vital component in the design and implementation of MIS and beyond.

3 Ontological Glossaries as Systems for Informational Documentation Management

In the next development of this project, we introduce SIDOC (System for Informational Documentation Management), a comprehensive tool designed for the creation and management of metadata information systems.

SIDOC represents each concept—derived from the domain expert's narrative—through three key elements: (i) a lemma, which names the concept, (ii) a definition, and (iii) the source of the definition. This approach is necessary because a lemma alone might not suffice due to the potential for homonymy, where different concepts share the same name. The tool's primary goal is to empower all project participants to verify whether the concepts they plan to use (for instance, in generating a report or conducting a statistical survey) have already been defined by others involved in the project. Ensuring consistent semantics for these concepts is crucial for the seamless integration of the various subgraphs that collectively form the entire project.

Beyond the definition of concepts, SIDOC plays a role in metadata documentation by elucidating the relationships between concepts. These relationships can vary, spanning from classic hierarchical structures typical of standard coding to the various verbs domain experts employ to articulate the discourse domain. A relationship of particular importance is "is measured by." By establishing this connection, each concept is linked with the statistical measures pertinent to its measurement. This linkage is useful for bridging the divide between a concept's qualitative descriptive use and its quantitative description.

3.1 Ontological Glossaries in Support of the Dominion Experts' Description

The Fig. 1 illustrates the application of the MIS_SIDOC system in a project focused on identifying the skills of migrants upon their arrival in Italy. In the upper portion of the image, a network is displayed, showcasing graphs generated from descriptions provided by domain experts. Each node within these graphs represents a concept, accompanied by its definition, and the connections between nodes signify the predicates. Noteworthy in the visualization is the use of various shapes; for example, hexagons indicate broader categories to which the nodes belong. Moreover, different colors are used to denote distinct contexts, such as demographics and work, allowing for nodes to be associated with multiple categories.

The lower part of the Fig. 1, linked to the upper by arrows, provides a detailed description of a specific node. This bifurcation demonstrates the system's capability to present information in a structured and interconnected manner.

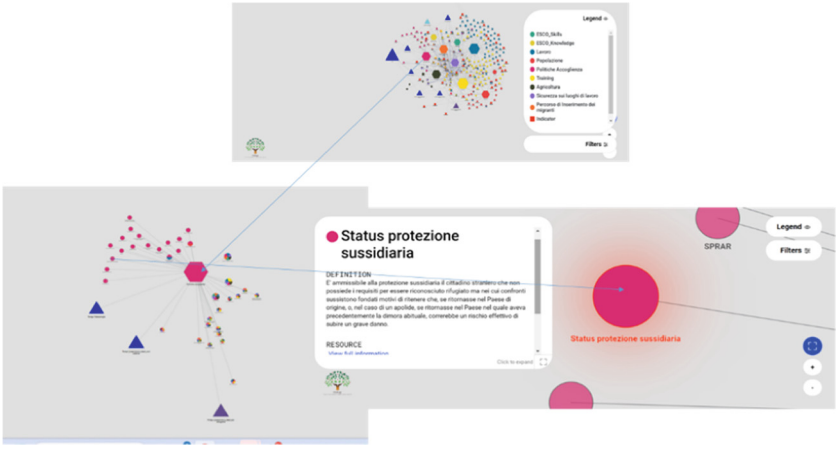


Fig. 1. SIDOC metadata network.

The system is equipped with comprehensive search functionalities, enabling users to locate the desired metadata. These include searches by keywords, attributes, relationships, and taxonomies, as well as full-text and semantic search options.

From an accessibility standpoint, MIS_SIDOC offers different access modes for its metadata: a web-based interface, a client-server user interface, or APIs.

3.2 Ontological Glossaries in Support of Integrable and Evolutive Information Systems

SIDOC supports the dissemination of the documentation, which is crucial for the automatic construction of conceptually coherent databases. After the metadata information system is generated based on descriptions from domain experts, SIDOC facilitates the creation of an ontology. This ontology, formulated in first-order logic, enables the automatic generation of a relational database. Such a database is automatically constructed from the conceptual model derived from the narratives of domain experts and is utilized to store data necessary for depicting the modeled reality.

4 Conclusions and Further Development

After illustrating how the structure of graphs underpins all phases of constructing a statistical source that meets the evolving and integrating descriptive needs of reality, the article presented the example of SIDOC, an information system designed for managing metadata. SIDOC, starting from the network of concepts used in describing and problematizing the field of investigation, not only facilitates the management of meta-information but

also automates the generation of relational databases necessary for physically archiving data.

The potential for interesting advances arises through the integration of the graphs generated by SIDOC with Large Language Models (LLMs) [9, 10]. This integration represents a promising new approach for generating relational databases. Specifically, LLMs leverage their learning and text-generation capabilities to automate schema creation and data insertion. This synergy offers an innovative method for database generation. Several ongoing research projects are actively exploring the potential of this integration in this field.

References

1. Teorey, T.J., et al.: Database Modeling and Design: Logical Design. Elsevier (2011)
2. Wand, Y., et al.: Theoretical foundations for conceptual modelling in information systems development. Decis. Support. Syst. **15**(4), 285–304 (1995)
3. Rensink, A.: Representing first-order logic using graphs. In: International Conference on Graph Transformation. Springer, Heidelberg (2004)
4. Dau, F.: Concept graphs and predicate logic. In: International Conference on Conceptual Structures. Springer, Heidelberg (2001)
5. Pavlov, V., Schukin, A., Cherkasova, T.: Exploring automated reasoning in first-order logic: tools, techniques and application areas. In: Knowledge Engineering and the Semantic Web: 4th International Conference, KESW 2013, St. Petersburg, Russia, 7–9 October 2013. Proceedings 4. Springer, Heidelberg (2013)
6. Bernstein, A., Halevy, A., Noy, N.: The challenges of semantic integration. IEEE Data Eng. Bull. **32**(2), 4–12 (2009)
7. Jeffery, K.G.: Metadata: the future of information systems. In: Information Systems Engineering: State of the Art and Research Themes. Springer, London (2000)
8. Schuurman, N., Leszczynski, A.: Ontology-based metadata. Trans. GIS **10**(5), 709–726 (2006)
9. Wang, C., Liu, X., Song, D.: Language models are open knowledge graphs. arXiv preprint arXiv:2010.11967 (2020)
10. Pan, S., et al.: Unifying large language models and knowledge graphs: a roadmap. IEEE Trans. Knowl. Data Eng. (2024)

Clickstream Data Analysis and Web User Profiling via Mixture Hidden Markov Models

Furio Urso(✉), Antonino Abbruzzo, Marcello Chiodi, and Maria Francesca Cracolici

Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy
furio.urso@unipa.it

Abstract. Using Mixture Hidden Markov Models (MHMMs), the study analyses clickstream data to identify users' profiles with similar browsing behaviour. MHMMs enable us to analyse categorical sequences, assuming they evolve according to a mixture of latent Markov processes, each related to a different subpopulation. An empirical analysis of clickstream data from a hospitality industry website has been performed. Evidence shows the usefulness of MHMMs in exploring user behaviour and defining ad-hoc marketing strategies. Finally, as MHMMs entail identifying two latent classes, viz., the number of sub-populations and hidden states, the study proposes a model selection criterion based on an integrated completed likelihood approach that accounts for both latent classes.

Keywords: customer behaviour · digital devices · clustering · hidden Markov models · entropy measure · model selection

1 Introduction

Clickstream data serve as crucial sources of information for businesses seeking insights into users' activity on their websites [2]. However, a major limitation of clickstream data lies in the lack of information regarding the reasons behind users' navigational goals. Extracting such information from the data, which allows for variations in browsing behaviour and the identification of distinct subpopulations, would prove immensely valuable for companies seeking to identify user profiles and define ad-hoc marketing strategies. Mixture Hidden Markov Models (MHMMs) are used as a model-based clustering approach to identify such user profiles. Specifically, users' visits to the website are represented by sequences of selected web pages; the sequences that belong to the same cluster evolve according to a hidden Markov process and the hidden states represent users' changes in goals and "attitudes" during exploration.

Previous research has explored the utilization of Mixture Hidden Markov Models (MHMMs) in the analysis of web browsing behaviour. For example,

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 230–235, 2025.
https://doi.org/10.1007/978-3-031-64346-0_39

[6] proposed the application of MHMMs as a clustering technique to categorize diverse sequences of observations, demonstrating their versatility in understanding web browsing behaviour and recognizing their potential for uncovering nuanced user interactions with online platforms. Building upon this foundation, [5] extended MHMM-based approaches to analyze web sequences across various browsing sessions for individual users, highlighting the adaptability of these models to different contexts. Furthermore, [9] applied an MHMM to automatically classify users and identify new web page categories via hidden states.

Specifically, our study addresses the challenge of selecting the number of mixture components and hidden states in MHMMs, drawing on established Information Criteria (ICs) like BIC and AIC. To overcome limitations, we introduce BIC_H, a novel entropy-based model selection criterion tailored for MHMMs, inspired by [1]'s ICL_BIC and [8]'s mixed HMM modification. BIC_H offers robust clustering for user profiling as it identifies users' groups, accounting for the hidden states. By applying BIC_H to clickstream data from LovePanormus, we identify user profiles based on browsing behaviour, enriching empirical research. This study applies MHMM to analyze clickstream data collected from the website of LovePanormus, a company operating in the hospitality sector, and identifies user profiles by proposing an entropy-based model selection criterion that accounts for the two latent classes in the model: clusters and hidden states. The paper is structured as follows: Sect. 2 introduces Clickstream data; in Sect. 3, MHMMs and the new entropy-based criterion are presented. Section 4 illustrates the case study of LovePanormus and the concluding remarks.

2 Clickstream Data

Clickstream data, which capture users' activity on a website, provides insights into their browsing behaviour by recording the web resources they access, such as web pages, images, and links, along with associated information like IP addresses, devices, browsers, and software used for access. To analyze clickstream data effectively, cleaning and processing steps are necessary to extract relevant information and create a sequential list of access requests that represent users' movements [4]. Specifically, in a first step, data are cleaned by removing web resources that are not the main focus of the analysis (e.g. images, links), obtaining a dataset of visited web pages. Then, data are processed to extract new information, such as users' geographical locations, from their IP addresses.

One of the primary challenges with clickstream data is the anonymity of user interactions, as IP addresses are reassigned over time, potentially leading to multiple users being associated with the same IP. While user registration or acceptance of cookies can mitigate this issue, many users may prefer anonymous browsing, in which case identification processes have to be implemented by exploiting information on devices and browsers or information related to website structure.[1] Another critical issue consists in identifying and filtering out *Bots*,

[1] For example, an IP address can identify two users if it selects a resource that is not directly accessible from the previous resource.

i.e., automated programs that mimic user behaviour. Various methods, such as monitoring access patterns and detecting rapid changes in devices or browsers, can be used to identify and eliminate *Bots*. In the next section, MHMMs will be briefly introduced, as well as a proposal of a selection criterion for the number of profiles based on an integrated completed likelihood approach.

3 Methods: MHMM and BIC_H

Mixture hidden Markov models (MHMMs) [7] cluster web sequences, assuming each mixture component (i.e., the k-th HMM) generates a set of sequences representing specific browsing behaviour, characterized by parameters $\Theta^k = \{\pi^k, A^k, B^k\}$ where π^k are initial probabilities, A^k are transition probabilities between hidden states, and B^k are emission probabilities between hidden states and observed states (web pages). The n web sequences y_i, $i = 1, 2, \ldots, n$, have length T with $\mathrm{card}|y_{it}| = R$ web pages, and u_i are hidden sequences such that $\mathrm{card}|u_{it}^k| = S^k$ hidden states in cluster k. The MHMM log-likelihood is:

$$\ell(\Theta, \omega; y, X) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_{ik} \sum_{u} \pi_{u_{i1}}^k b_{u_{i1}}^k(y_{i1}) \prod_{t=2}^{T} a_{u_{i,t-1},u_{it}}^k b_{u_{it}}^k(y_{it}) \right) \quad (1)$$

Sequences are assigned to a component by updating membership $\omega_{ik} = P(M^k|X_i)$, depending on covariates X_i (e.g., visit-related information like access device, IP nationality, etc.)

To select the number of cluster K and hidden states for each cluster $S = \{S^1, S^2, \ldots, S^K\}$, a selection criterion based on an approximation of the Integrated Completed Likelihood is defined based on the ICL_BIC [1]. The new criterion for MHMM is called BIC_H and is a entropy-penalized BIC where the penalization term is a sum of entropies for each sequences y_i, $i = 1, 2, \ldots, n$.

$$\mathrm{BIC}_H = \ell(\Theta, \omega; y, X) - \sum_{i}^{n} \mathrm{H}(M, U|y_i, X_i) - (\log N)\frac{\mathrm{df}}{2} \quad (2)$$

where $N = n \times T$, df are model degrees of freedom, and $\mathrm{H}(M, U|y_i, X_i)$ is the joint entropy for sequence y_i, obtained as the sum of two elements: the entropy related to clusters and the entropy related to hidden sequences.

$$\begin{aligned}
\mathrm{H}(M, U|y_i, X_i) &= \mathrm{H}(M|y_i, X_i) + \mathrm{H}(U|M, y_i, X_i) \\
&= -\sum_{k=1}^{K} P(M^k|y_i, X_i, \hat{\Theta}) \log P(M^k|y_i, X_i, \hat{\Theta}) + \\
&\quad -\sum_{k=1}^{K} P(M^k|y_i, X_i, \hat{\Theta}) \left[\mathrm{H}(U_{i1}|y_i, M^k, \hat{\Theta}) + \sum_{t=2}^{T} \mathrm{H}(U_{it}|U_{i,t-1}, y_i, M^k, \hat{\Theta}) \right].
\end{aligned}$$

where the initial and transitional entropies in the third equation are obtained as proposed by [3]. In the next section, we present an application of this approach to real clickstream data.

4 Results and Conclusion

MHMM was used to identify behavioural profiles by analyzing clickstream data from a Sicilian holiday-provider, LovePanormus.it. The original dataset comprised 2,487,802 observations, representing web resources in chronological order from September to December 2017. Data were cleaned by removing irrelevant log lines and suspected bots. The refined dataset contained only HTML pages, totaling 95,201 lines. User browser, software, and device information were extracted using the *uaparserjs* R package. Session identification relied on a time-based approach due to anonymous access, with a 10-minute threshold to distinguish sessions. Sessions with only three clicks were removed, resulting in 10,252 user sessions. Sequence elements are page categories (i.e., website thematic areas) instead of actual web pages, due to the number of pages and the speed with which pages were changed in the referred period. These categories are: the "Homepage", the "Attractions" area, the "Accommodation" area, the "Events" area, the "Experiences" area, the "Services" area, and the "Info" area about the company staff and partners. The model selected was the one minimizing the criterion BIC_H, specifically one with $K = 3$ clusters/profiles having numbers of states $S = 3, 2, 4$. The three identified clusters have been labelled as follows: *i)* Casual Explorers and Potential Partners; *ii)* Information Seekers; and *iii)* Potential Tourists.

The profile of *Casual Explorers and Potential Partners* includes 22% of the users, mostly desktop accesses. The three hidden states in this cluster represent three different "attitudes" that users adopt while exploring the website. One is related to a lack of interest in the website, mostly characterized by selecting "Homepage" area; the second is related to an exploratory behavior; and the third is related to users that explore "Info" area.

As concerns the profile of *Information Seekers,* it includes 19% of the users; desktop access is the preferred one, but users in this cluster have the highest percentage of mobile use (38% of the cluster). There are two hidden states, representing two "attitudes": a focus on the "Attractions" pages or a more diversified exploration with a preference for accessing the "Events" area.

Finally, the profile of *Potential Tourists* consists of 59% of users that prefer to access via desktop. There are 4 hidden states representing: an initial phase in the website "Homepage" area, then users move to an exploratory "attitudes" that focus on "Attractions" area or a purchase-oriented one that focuses on "Accommodation" area. There is also a fourth "attitudes" related to accessing "Experiences" area that is rarely selected by users in this cluster.

We summarize the three clusters' structure in Figs. 1, 2 and 3 as a directed graph that represents user behaviour. Pies are the hidden states, edges are transitions between states, and pie slices are the emission probabilities that connect hidden and observed states (thematic areas).

Our analysis indicates that the website caters to two primary user segments, each associated with distinct business models. The last two clusters primarily

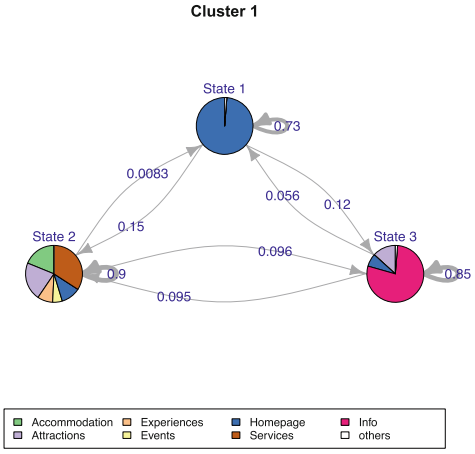


Fig. 1. Casual Explorers and Potential Partners

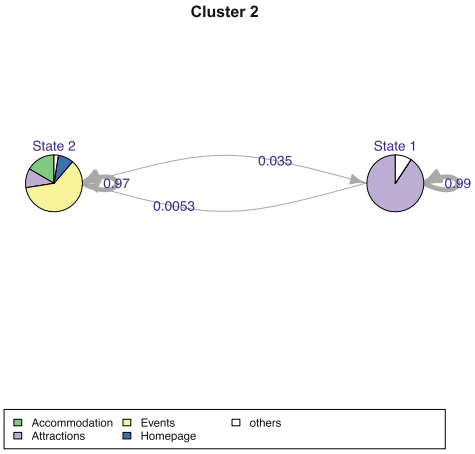


Fig. 2. Information Seekers

attract potential tourists interested in exploring or purchasing tourist-related products and services, reflecting the firm's adoption of a business-to-consumer (B2C) model. On the other hand, Cluster 1 likely comprises competitors interested in checking the company information. This suggests that LovePanormus may need to reconsider the scope of its B2C model and explore opportunities for business-to-business (B2B) sales to diversify its market and revenues.

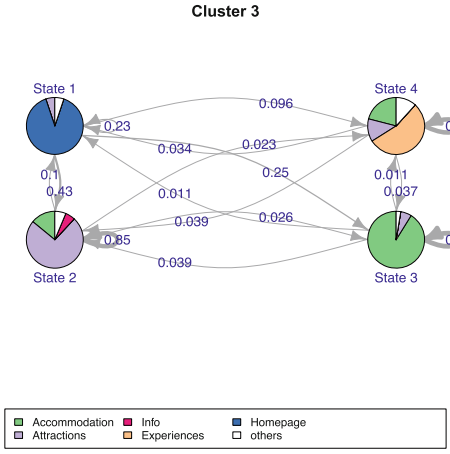


Fig. 3. Potential Tourists

References

1. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**(7), 719–725 (2000)
2. Das, R., Turkoglu, I.: Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Syst. Appl. **36**(3), 6635–6644 (2009)
3. Durand, J.B., Guédon, Y.: Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models. Stat. Comput. **26**(1–2), 549–567 (2016)
4. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. 1. Springer, Berlin (2011)
5. Scott, S.L., Hann, I.H.: A nested hidden Markov model for internet browsing behavior. Marshall School of Business, pp. 1–26 (2006)
6. Smyth, P.: Probabilistic model-based clustering of multivariate and sequential data. In: Proceedings of the Seventh International Workshop on AI and Statistics, Citeseer, pp. 299–304 (1999)
7. Vermunt, J.K., Tran, B., Magidson, J.: Latent class models in longitudinal research. Handbook of longitudinal research: Design, measurement, and analysis, pp. 373–385 (2008)
8. Volant, S., Berard, C., Martin-Magniette, M.L.: Hidden Markov models with mixtures as emission distributions. Stat. Comput. **24**(4), 493–504 (2014)
9. Ypma, A., Heskes, T.: Automatic categorization of web pages and user clustering with mixtures of hidden Markov models. In: International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles, Springer, pp. 35–49 (2002)

Analysing Language for Preventing Women from Gender Violence: NLP and Machine Learning Techniques to Classify Tweet Messages

Fiorenza Deriu[1(✉)] and Emilia La Nave[2]

[1] Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy
fiorenza.deriu@uniroma1.it

[2] Institute for Complex Systems, National Research Council, Piazzale Aldo Moro 7, 00185 Rome, Italy
emilia.lanave@cnr.it

Abstract. Gender violence is often overlooked, with few women seeking help from anti-violence shelters. This study explores how to effectively recognise women at risk of violence through the combined use of multidimensional statistical techniques such as text mining, Natural Language Processing, and machine learning to detect signs of violence in social media posts. Analysing the characteristics of the language used by social media users, this research aims to develop an automated system for the early detection of women potentially exposed to gender violence. This information could be employed to provide these selected social media users with details on available healthcare, psychological, and legal support services.

Keywords: gender violence prevention · machine learning · text mining · NPL

1 Introduction

Gender violence is still largely invisible to society and statistics. In accordance with the latest available Istat data a relevant number of women do not consider violence a crime [1], and in fact only the 35.4% of those who experienced physical or sexual violence by a partner believe being victim of a crime. Many women neither talk to anyone about the violence they suffer from their partner (28.1%), nor report to police officers (reporting rate 12.1% when the author is the partner - 6% if non-partner). There are still a limited number of women seeking help from anti-violence shelters, or from other specialised services (about 4% violence inside the relationship - 1% outside). In most of the cases women do not know that these services exist. So, despite the increasing attention given to the issue, many women are unaware that they are victims of violence and do not seek support and assistance from services. Therefore, the research question guiding this paper is: is there

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 236–241, 2025.
https://doi.org/10.1007/978-3-031-64346-0_40

an automatic way to recognise and deliver messages and information to women potentially at risk of violence that can help them become aware of the situation they are experiencing, thus encouraging them to seek assistance from specialised centers in the area?

The presented study aims to identify a procedure that allows for the reliable recognition of contents related to violence against women in its various manifestations (sexual harassment at work, stalking, rape, physical and sexual violence, etc.), within generic texts posted on social network platforms. The hypothesis underlying this research is that the recognition of posts containing such contents may provide valuable insight on the potential exposure of speakers to gender violence. Social media platforms, in fact, represent a space within which women, by participating in social discussions on the topic of violence, albeit without referring to themselves, may leave traces of their experience. Techniques such as Text Mining [2], Natural Language Processing (NLP) [3], and the application of multidimensional statistical methodologies and machine learning to texts written by women who might be (or not) victims of violence could allow for the identification of the characteristic language used by potential victims. This knowledge could be used to develop automatic mechanisms for recognising and classifying at-risk individuals, enabling the early identification of potential victims and directing them to health, psychological, and legal support services. This approach has already been successfully applied in attempting to predict and characterise similar social and health issues such as suicide [4], bullying [5], and domestic violence [6,7]. We suggest that, similarly to what already the commercial recommendation systems do, once a message is classified as concerning gender violence, the platform may start sending to those post or messages information that may encourage the receiver to get in touch with specific services or national anti-violence public service numbers. Therefore, the use of the proposed method may support preventative actions to combat violence against women.

2 Data and Methods

In order to achieve the aforementioned objectives, a random sample of 20,992 italian tweets, collected in 2019, has been analysed. The tweets were selected according to the following 11 hashtag related to violence on women: #quellavoltache, #femminicidio, #nonunadimeno, #stupor, #violenzacontroledonne, #violenzegenere, #violenzadonne, #violenzadulledonne, #molestie, #molestiesessuali, and #sessismo. The corpus was originally made up of 425,823 token (N), 30.317 types (V), and 16,466 hapax. The mean of occurrences per text was equal to 20.29. At this stage, using both Iramuteq software and Phyton Libraries NLTK and SPACY, the corpus was pre-processed in order to remove punctuations, urls, mentions, and emoticons as well as the hashtag used for tweet's selection, the hapax, and the stop-words. The pre-processed corpus was lexicalised, after identifying the most relevant n-grams, repeated segments (n = 2, 3, and 4), and the most common multi-words. At the end of the first stage process the corpus had 395,653 token (N), 40,434 types (V), and 21,725 hapax. The

Table 1. Lexicometric indicators of the corpus at the end of the first stage process.

Indicator	Value
Type Token Ratio	$TTR = V7N * 100 = 10.2 < 20\%$
Guiraud index	$Guiraud = V/\sqrt{N} = 64.2 > 22$
Zipf law	$Zipf = logN/logV = 1.2. \rightarrow \approx 1.3$
Hapax %	$Hapax = V1/V = 53,7\%$

mean of occurrences per text was equal to 9.8. Hence, lexicometric indicators have been computed on the corpus to validate its suitability for Natural Language Processing techniques (Table 1). Finally, the analysis of specificities and similarities (co-occurrencies) has enabled the identification of the main topics characterizing the tweets contents. In total, the following 6 main thematic areas were identified: individuals, contexts, feelings, protest actions and claims, types of violence against women, and symbolisms associated with gender discrimination and violence against women. Then, in order to identify the tweets to use at the second stage of the study, to train the machine learning algorithm, a cluster analysis on text segments based on the classification method of Reinert[1] has been carried out. The cluster analysis has allowed to identify 4 clusters of words related to the following themes: gender unbalance, discrimination, sexism, and machism (cluster 1); sexual harassment at work and the Weinstein case (cluster 2); sexual violence, rape, femicide, and judicial verdicts (cluster 3); and finally collective protests and advocacy by feminist movements (cluster 4) (Fig. 1). The tweets classified in the first three clusters, totalling 16,641, due to their thematic specificity, were used in the second stage to train the machine learning algorithm to build a classifier that allows us to automatically identify tweets using specific language related to gender-based violence. To this end, the process flow shown in Fig. 2 was adopted. The corpus used for training the classifiers consists of 73,346 tweets. Out of these, 16,641 tweets belong to the gender-based violence cluster selected at the first stage of the study, and 56,705 pertain to other topics. To train the classifiers, the total corpus has been split into two subsets (Fig. 2): two-thirds of the available tweets were randomly selected from the initial set and used for training the classifiers, while the remaining 33% of the available texts, amounting to 24,424 tweets, was used to evaluate the predictive capabilities of the classifier. After pre-processing and normalizing the data, considering multiwords according to the criteria adopted at stage 1, tools offered by the Scikitlearn (sklearn) module for machine learning in the Python programming language were used for the analysis. For the reduction to graphical forms and the necessary transformation into a matrix form (feature x document) for the analysis, two different tools available in sklearn were employed: bag-of-words (CountVectorizer) and TfidfVectorizer. In the latter case, terms transformed into vector forms are weighted by the Term Frequency-Inverse Document Frequency

[1] The method of Reinert is based on a descendant hierarchical classification algorithm.

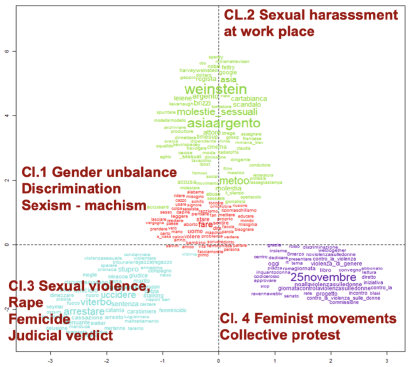


Fig. 1. Cluster analysis of Reinert.

Fig. 2. Process flow for classifier selection.

(TF-IDF) factor, according to a method that recognizes as more significant words that occur more frequently in only a few documents. The classifiers compared in this initial phase of the study are those commonly used for binary classification: the Multinomial Naive Bayes model (NB), the Logistic Regression Classifier (LR), the Support Vector Machine classifier (SVM), the Stochastic Gradient Descent Classifier (SGD), the Multi-layer Perceptron Classifier (MLP), and the RandomForest Classifier (RF). The CountVectorizer or TfidfVectorizer methods are tested on all the classifier to compare their efficiency. The parameters of the classifiers are optimized using the Grid Search Cross Validation method.

3 Results

To evaluate the performance of the classifier, the confusion matrix has been built up. It is a 2×2 matrix where the elements represent: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). The model has been evaluated according to the following indicators: a) the accuracy, computed as the ratio of the sum of correct results (TP + TN) divided by the sum of all values (TN + FP + FN + TP); b) the value of the F score, which measures the weighted harmonic mean of Precision P (the ratio of TP to the sum of TP and FP) and Sensitivity S (the ratio of TP to the sum of TP and FN) (F score = 2SP/(S+P)); c) the AUC value, corresponding to the area underlying the ROC curve, the plot of the true positive rate against the false positive rate at different threshold values. Based on the analysis of the indicators, it was possible to conclude that, with the exception of the Naive-Bayes one, all classifiers showed good performance, with an accuracy ranging from 0.89 to 0.91 and an F score from 0.70 and 0.78, as showns in Table 2. SVM classifier scored the best AUC value. The confusion matrix and the ROC curve for SVM classifier are shown in Fig. 3. In conclusion, the results demonstrate that it is possible to use machine

Table 2. Comparison of performance indicators.

	Tf-idf Vectorizer						Count Vectorizer					
	NB	SVM	LR	SGD	RF	MLP	NB	SVM	LR	SGD	RF	MLP
Accuracy	0.81	0.90	0.89	0.88	0.91	0.91	0.83	0.89	0.90	0.90	0.91	0.90
Fscore	0.43	0.76	0.73	0.70	0.77	0.77	0.42	0.76	0.76	0.75	0.78	0.78
AUC	0.55	0.94	0.74	0.71	0.78	0.77	0.50	0.94	0.71	0.73	0.78	0.77

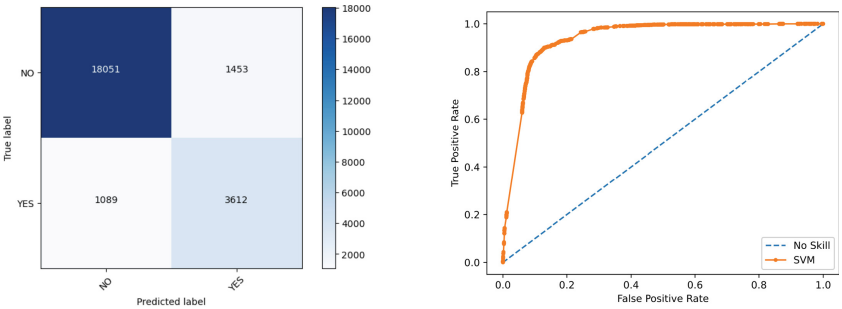


Fig. 3. Confusion Matrix (a) and ROC curve (b) for SVM Classifier.

learning techniques to recognise tweets concerning gender-based violence with a considerable level of accuracy. However, the analysis performed on the significant keywords for the classifiers suggests that only a part of the vast and nuanced lexical world of gender-based violence is captured, and that the corpus studied in this article represents a circumscribed subgroup. For greater completeness of the results, it would be important to extend the analysis to corpora based on social media post capable of capturing the moods of potential victims [7].

4 Conclusions

One of the major obstacles in building efficient language classification algorithms lies on the difficulty of constructing large-scale database containing high-quality information required to train the classifiers. This difficulty increases further when the language deals with sensitive topics such as violence against women. Furthermore, constructing vocabularies that account for the complexity of the Italian language presents an even greater challenge. In this article, we have proposed an innovative forward and backward method combining different techniques aimed at identifying well-focused topics, that could improve the capacity of machine learning algorithms for automatic detection of posts and messages concerning specific topics of interest. The ultimate aim of the research lies in the development of algorithms for the recognition and automatic classification of posts published by women potentially exposed to a variety of gender-based violence

abuses. This research may have a significant positive impact on society. In fact, the early detection of a significant number of potential victims of gender violence, who might otherwise remain unnoticed, could be used to support preventative measures. The study suggests that the early detection of potential situations of violence may be possible analysing the language used on social media. In fact, the tweets related to gender-based violence can efficiently be detected in an automated way, even if it does not mean that those who are talking about gender violence are victim of such a violence. For this reason, it would be worth continuing this research line, including in the analysis tweets involving the emotional sphere, and that could be extracted from thematic forums, Facebook groups, Instagram, TikTok, or blogs. Moreover, the adoption of an interdisciplinary approach, integrating linguists, sociologists, and data scientists, would be strongly recommended.

Aknowledgement. This work has been developed with the financial support of the Italian Ministry of University and Research within the 2022 PRIN program - Grant. Z85NCT (NextGenerationEU - M4.C2.1.1 - CUP B53D23019750006).

References

1. Istat, Indagine sulla Sicurezza delle donne. https://www.istat.it/it/archivio/5348 (2016)
2. Bolasco, S.: L'analisi automatica dei testi. Carocci Editore, Roma (2013)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python, O'Relly Media (2009)
4. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual Association for Computing Machinery Web Science Conference, WebSci 2013, pp. 47-56, ACL New York, NY, USA (2013)
5. Xu, J.M., Burchfiel, B., Zhu, X., Bellmore, A.: An examination of regret in bullying tweets. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 697–702, ACL Atlanta, Georgia (2013)
6. Schrading, N.: Analyzing Domestic Abuse using Natural Language Processing on Social Media Data (Thesis) Rochester Institute of Technology RIT Scholar (2015)
7. Subramani, S., Quan Vu, H., Wang, H.: Intent classification using feature sets for domestic violence discourse on social media, arXiv:1804.03497v1 (2018)

A Pogit Model for Under-Reported Counts of Violence Against Women in Italy

Silvia Polettini[1(✉)], Sara Martino[2], and Greta Panunzi[3,4]

[1] Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, P.le Aldo Moro, 5, Rome, Italy
silvia.polettini@uniroma1.it

[2] Department of Mathematical Science, Norwegian University of Science and Technology, Alfred Getz' vei 1, Trondheim, Norway
sara.martino@ntnu.no

[3] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le Aldo Moro, 5, Rome, Italy
greta.panunzi@uniroma1.it

[4] Dipartimento di Scienze Umane e Sociali, Università del Salento, Lecce, Italy

Abstract. Violence against women (VaW) is difficult to quantify as it is typically a largely underrated phenomenon. Clearly, gender-based violence could be more effectively contrasted if policies were informed by up-to-date and comprehensive evidence. Sample surveys are acknowledged to be the most reliable and established method to estimate the prevalence of violence and its characteristics, but when recent specialized survey data are not available (the last survey in Italy dates back to 2014), administrative data sources such as police registers can be considered as a possible source of information, though affected by large underreporting. We propose to model these data by a Poisson regression, explicitly accounting for the under-reporting, using the so-called Pogit model. To inform our model we include the available information on both the reporting process and the event intensity obtained from additional data sources such as the 1522 helpline number database and the BES system of indicators.

Keywords: Bayesian hierarchical model · compound Poisson · Pogit model · underreporting · violence against women

1 Introduction

Violence against women (VaW) is difficult to quantify as it is a largely underrated phenomenon; at the same time, a high level of underreporting contributes to protect and perpetrate the violent environments.

Supported by 2022 PRIN Grant Z85NCT (NextGenerationEU - M4.C2.1.1 - CUP B53D23019750006) - Italian Ministry of University.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 242–247, 2025.
https://doi.org/10.1007/978-3-031-64346-0_41

Official figures such as police statistics and justice registers are affected by large underreporting. Clearly, efficient and effective planning of policies to contrast gender-based violence dramatically depends on availability of up-to-date and comprehensive evidence.

To assess the prevalence and the characteristics of the phenomenon, many countries conduct specific sample surveys, as they are acknowledged to be the most reliable and established method to this aim. Such surveys are clearly complex and expensive to carry. In Italy, they have recently been included as an integral part of the national data collection system, but the last survey dates back to 2014. In the absence of recent ad-hoc survey data, administrative data sources such as police registers may be used as an alternative source of information, aware of the large underreporting that affects such data. From a statistical perspective, in order to assess the phenomenon it is important to exploit all the available information [10,19], taking into account the limitations and issues of each data source.

We consider a Poisson regression model on official police reports that explicitly accounts for the under-reporting mechanism, and use the available information to inform both the reporting process and the event intensity to produce prevalence estimates.

2 Key Socio-Demographic Correlates of Violence and Underreporting

As confirmed in the literature [3,5,9], VaW is significantly associated with social and cultural aspects, primarily exposure to violent environments, and exhibits variations across life stages.

Considering the determinants of violence, numerous studies have delved into the association between violence-supportive attitudes and the perpetration of violence against women, as highlighted in references like [1,7]. The perpetration of violence may be considered as a way to achieve women's subordination. Not only are men's sexist, misogynist, or patriarchal attitudes linked to the use and tolerance of violence against women, but women's adherence to conservative gender and sexuality norms also plays a fundamental role. The promotion of education and labor participation plays a crucial role in fostering the economic independence and social integration of women, and expectedly reduce the risk of victimization as well as the level of underreporting. The socio-economic factors of labor market participation and socioeconomic status have been identified as influential in shaping attitudes towards violence against women: [7] highlights associations between economic and social disadvantage and higher violence rates, attributing these to both violence-prone attitudes and higher exposure to violence. Additionally, educational attainment and age are recognized as potentially influencing individuals' perceptions of violence against women and, consequently, impacting their willingness to report violence episodes. Gender role norms and women's beliefs about gender roles and sexuality are also pivotal in shaping perceptions of experiences and responses to violence, influencing the likelihood

of disclosing such incidents. These relationships are, in turn, influenced by the continuously evolving cultural context.

Note that the higher reporting rates may lead to paradoxical results and mixed conclusions if underreporting is not appropriately accounted for (see, for example,[16] and references therein).

The level of underreporting may be ascertained by ad hoc survey data; the reporting rates are quite low in Italy as, according to the 2014 Italian Safety Survey, only a 12,2% of abuses suffered by partner and a 6% of abuses suffered by non-partner were reported to the police.

The database of calls to the helpline number 1522 data administered by Istat[1] also offers some information about reporting. Among victims reaching out to the helpline number, for which the operator could investigate for the user having reported to the police, the percentage of users that do report violence to the police is about 17% in 2020, and about 19% and 20% in 2019 and 2018 respectively, which confirms a low propensity to report to the police, especially in 2020.

Within this context, our focus lies on investigating the factors influencing violence and under-reporting, respectively, trying to use all the available information on both aspects to draw a picture of VaW in Italy.

3 The Model

Count data are usually analysed through Poisson models. The intensity of the event can further be expressed in terms of covariates, and additional spatially structured random effects may be added.

In order to represent the misreporting mechanism typical of our data, we include in the probabilistic model of the outcome an additional layer of uncertainty.

This approach allows us to combine the existing information from crime reports with other sources, with the aim of generating prevalence estimates even in the absence of dedicated survey data. Moreover, it allows for finer geographical detail compared to what is achievable through traditional surveys.

Let T_i denote the number of events of violence against women occurred over a set of m areas. Let

$$T_i \sim Poisson(E_i\theta_i), \qquad i = 1, \ldots, m$$

where θ_i is the event rate and E_i is the number of women in the i–th area. Let us further assume that we can relate the rates θ_i to a set of covariates $X_1, ..., X_p$ through a regression model such as:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}.$$

As the events of violence are only partially reported, we observe reported counts $Y_i, Y_i \leq T_i$, $i = 1, \ldots, m$.

[1] https://www.istat.it/it/archivio/278050.

The aim of the analysis is twofold, namely estimating the event intensities $\theta_i,\ i = 1, \dots m$ and the effects of the covariates, and predicting the true counts $T_i,\ i = 1, \dots m$ given the available information.

To account for under-reporting several proposals have been proposed, all relying on expert information and/or auxiliary data on the reporting process.

A vast literature, e.g. [6,12,15,17,18] uses Poisson stopped-sum distributions (see [11], Section 4.11) to model the observed counts. Under this approach, the severity, rather than the mere chances of under-reporting, can be estimated along with the true count prevalence.

The model that we describe next is a particular case within this class. Assuming that each individual h in area i has area-specific probability ϵ_i to report the event, independently, $h = 1, \dots, T_i$, $i = 1, \dots, m$, and independent on T_i, the observed counts can be written as

$$\begin{aligned} &Y_i = \sum_{l=1}^{T_i} W_{ih}, \quad W_{ih} \overset{ind}{\sim} \text{Bernoulli}(\epsilon_i), \quad h = 1, \dots, T_i \\ &Y_i \mid T_i, \epsilon_i \sim Bin(T_i, \epsilon_i), \\ &T_i \sim \text{Poi}(\theta_i). \end{aligned} \tag{1}$$

Marginalising over T_i one gets

$$Y_i \mid \theta_i, \epsilon_i \sim Poisson(E_i \theta_i \epsilon_i). \tag{2}$$

Moreover, for the number of non-reported events it holds that

$$T_i - Y_i \mid \theta_i, \epsilon_i \sim Poisson(E_i(1 - \theta_i)\epsilon_i). \tag{3}$$

[4,15] propose a Bayesian approach for count data using distribution (2) and including covariates to inform not only the true count-generating process but also the under-reporting mechanism; this approaches also allows for complex spatio-temporal structures. [15] model the reporting probability hierarchically through a logistic regression using suitable covariates, the so-called Pogit model. We consider the following hierarchical model:

$$Y_i \mid \theta_i, \epsilon_i \sim Poisson(E_i \theta_i \epsilon_i) \tag{4}$$

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + u_i + v_i \tag{5}$$

$$logit(\epsilon_i) = \gamma_0 + \gamma_1 Z_{1i} + \dots + \gamma_r Z_{ri} \tag{6}$$

u_i, v_i being unstructured and spatially structured random effect that accounts for unobserved risk factors and heterogeneity across the study areas, modelled via intrinsic CAR model as in the reparametrization of the BYM model [2] that has been suggested by [14].

An issue of identifiability arises for model (4)–(6): indeed, whereas the product $\theta_i \epsilon_i$ is identified from the observations, θ_i and ϵ_i are not, since the same likelihood is obtained for all the combinations of θ_i and ϵ_i that give the same value of $\theta_i \epsilon_i$. In non-identified models, the Bayesian approach is particularly appropriate in that, as long as the prior is proper, inferences are guaranteed because

the posterior distribution is well defined and therefore MCMC algorithms can be designed to simulate from the posterior (see [8] for a discussion of the issue of identification in the Bayesian vs frequentist approach). In the absence of any completely reported observations, the non-identifiability issue can be addressed by introducing external information on one or both of the models for θ_i and ϵ_i, as discussed in the literature. Informing the reporting process based on expert knowledge or external information ensures the convergence of MCMC algorithms to the posterior, see the discussion in [13,15] and references therein for details. As proposed in [15], it is sufficient to specify an informative prior on the intercept of the logit model for the reporting probability. This approach is advantageous as one can rely on past data or expert insights on the mean reporting rate. [4] suggest to elicit a prior on the reporting probability π_0 at the "average" value of all covariates, by defining a beta prior with prespecified mode and $q-$quantile. This implies a prior for the intercept γ_0 of the logit model (6). Except for the latter parameter, we specify vague normal distributions for the regression coefficients, namely $\beta_j \sim N(0, 10^2)$, $j = 0, \ldots, p$ and $\gamma_j \sim N(0, 10^2)$, $j = 1, \ldots, r$.

We apply the model to official reports on battering and sexual violence obtained from the register of crime statistics administered by the Italian Ministry of Interior, including the available information on both the reporting process and the event intensity obtained from additional data sources such as the 1522 helpline number database, the BES system of indicators, and other indicators available at Istat, with the aim of highlighting the factors influencing violence and under-reporting, and to obtain area-level estimates of violence in Italy in the absence of up-to-date survey data.

Acknowledgments. Work developed under the support of 2022 PRIN Grant Z85NCT (NextGenerationEU - M4.C2.1.1 - CUP B53D23019750006) - Italian Ministry of University.

References

1. Ambrosetti, E., Amara, N.A., Condon, S.: Gender-based violence in Egypt: analyzing impacts of political reforms, social, and demographic change. Violence Against Women **19**(3), 400–421 (2013). https://doi.org/10.1177/1077801213486329. PMID: 23676450
2. Besag, J., York, J., Mollié, A.: Bayesian image restoration with application in spatial statistics. Ann. Inst. Stat. Math. **43**, 1–20 (1991)
3. Capaldi, D.M., Knoble, N.B., Shortt, J.W., Kim, H.K.: A systematic review of risk factors for intimate partner violence. Partn. Abus. **3**(2), 231–280 (2012)
4. Chen, J., Song, J., Stamey, J.: A Bayesian hierarchical spatial model to correct for misreporting in count data: application to state-level COVID-19 data in the united states. Int. J. Environ. Res. Public Health **19**(6) (2022). https://doi.org/10.3390/ijerph19063327
5. Copp, J.E., Giordano, P.C., Longmore, M.A., Manning, W.D.: The development of attitudes toward intimate partner violence: an examination of key correlates among a sample of young adults. J. Interpers. Violence **34**(7), 1357–1387 (2019)

6. Dvorzak, M., Wagner, H.: Sparse Bayesian modelling of underreported count data. Stat. Model. **16**, 24 (2016)
7. Flood, M., Pease, B.: Factors influencing attitudes to violence against women. Trauma Violence Abuse **10**(2), 125–142 (2009). https://doi.org/10.1177/1524838009334131. PMID: 19383630
8. Florens, J.P., Simoni, A.: Revisiting identification concepts in Bayesian analysis. Ann. Econ. Stat. (144), 1–38 (2021). https://www.jstor.org/stable/10.15609/annaeconstat2009.144.0001
9. Garcia-Moreno, C., Jansen, H.A., Ellsberg, M., Heise, L., Watts, C.H., et al.: Prevalence of intimate partner violence: findings from the who multi-country study on women's health and domestic violence. The Lancet **368**(9543), 1260–1269 (2006)
10. Gelles, R.: Estimating the incidence and prevalence of violence against women: national data systems and sources. Violence Against Women **6**(7), 784–804 (2000). https://doi.org/10.1177/1077801200006007005
11. Johnson, N., Kotz, S., Kemp, A.: Univariate Discrete Distributions. Wiley Series in Probability and Statistics, Wiley, Hoboken (2005)
12. Li, T., Trivedi, P., Guo, J.: Modeling response bias in count: a structural approach with an application to the national crime victimization survey data. Sociol. Methods Res. **31**(4), 514–544 (2003). https://doi.org/10.1177/0049124103251951
13. de Oliveira, G.L., Argiento, R., Loschi, R.H., Assunção, R.M., Ruggeri, F., D'Elia Branco, M.: Bias correction in clustered underreported data. Bayesian Anal. **17**(1), 95–126 (2022)
14. Riebler, A., Sørbye, S.H., Simpson, D., Rue, H.: An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat. Methods Med. Res. **25**(4), 1145–1165 (2016). https://doi.org/10.1177/0962280216660421. PMID: 27566770
15. Stoner, O., Economou, T., da Silva, G.D.M.: A hierarchical framework for correcting under-reporting in count data. J. Am. Stat. Assoc. **114**(528), 1481–1492 (2019). https://doi.org/10.1080/01621459.2019.1573732
16. Whaley, R.B.: The paradoxical relationship between gender inequality and rape: toward a refined theory. Gender Soc. **15**(4), 531–555 (2001). http://www.jstor.org/stable/3081921
17. Whittemore, A.S., Gong, G.: Poisson regression with misclassified counts: application to cervical cancer mortality rates. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **40**(1), 81–93 (1991). http://www.jstor.org/stable/2347906
18. Winkelmann, R.: Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. Empirical Economics **21**(4), 575–587 (1996). https://doi.org/10.1007/BF01180702
19. Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A., Blom, A.G.: Integrating probability and nonprobability samples for survey inference. J. Surv. Stat. Methodol. **8**(1), 120–147 (2020)

A Dynamic Spatial Indicator of the Surface Water Quality

Antonella Congedi[1,2,3(✉)] and Sandra De Iaco[1,2,3]

[1] University of Salento, Lecce, Italy
{antonella.congedi,sandra.deiaco}@unisalento.it
[2] National Centre for HPC, Big Data and Quantum Computing, Bologna, Italy
[3] National Biodiversity Future Center, Palermo, Italy

Abstract. The evaluation of water quality usually considers the effect of different physical and chemical parameters, which can vary across space and time. In this context, it is crucial to build a spatio-temporal composite indicator which might provide useful insights for a sustainable development and might support strategies for saving water and controlling pollution. For this aim, an advanced Geographically weighted principal components analysis indexed in time, based on the use of geostatistical tools, is proposed. Thus, this technique is applied on parameters observed for ten years, for the superficial rivers of Emilia-Romagna Region, which presents a complex water system.

Keywords: Water quality · GWPCA · Structural analysis · Temporal analysis

1 Introduction

In the literature, there can be found different contributions regarding the evaluation of water quality throught indexes, which consider several parameters in the computation, such as the National Sanitation Foundation Water Quality Index (NSF WQI) [8,12], Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI) [2], and Oregon Water Quality Index [1]. Since these indexes were introduced with different purposes, using the same parameters with different methods can lead to different results and water classifications. Despite their differences, the goal of all indexes is the same, summarise information in a single value, but whithout considering that parameters can be correlated in time and space.

In the evaluation of surface water quality, many physical and chemical parameters are controlled over time. For the characteristics of the phenomenon, all the observed variables can be correlated, thus multivariate methods have a key role. Among traditional methods, principal components analysis (PCA) has been already largely used to build synthetic indexes from numerous variables [3,4]. However, in the presence of multiple variables measured at several spatial and temporal points, the spatial and temporal correlation of observed data cannot

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 248–252, 2025.
https://doi.org/10.1007/978-3-031-64346-0_42

be ignored. In the spatial context, the traditional PCA can be replaced by the geographically weighted principal components analysis (GWPCA) which can be considered as geographically weighted regression combined with the PCA. Thus, GWPCA can be seen a localized PCA, which can reduce the dimensionality of the multivariate dataset, taking into account the spatial dependence. However, in this study, the temporal dimension is also included and a GWPCA indexed in time has been proposed. As a consequence, in the present contribution, a new water quality index (WQI) which considers the main feature of a multivariate spatial phenomenon over time has been proposed. The spatial monthly data measured, over 10 years (2010–2019) at 53 water monitoring sites, belonging to the Emilia-Romagna Region in the northern part of Italy, have been analyzed. After a brief theoretical framework (2), a case study concerning spatial water data collected over the Emilia-Romagna Region.

2 Theoretical Framework

In environmental studies, GWPCA [5,9] represents one of the techniques which merges the principal components analysis and the geographically weighted regression [6] in order to take into account the spatial profile of a multivariate data set. However, when the variables are also controlled over time, it is often convenient to introduce the temporal dimension (other than the spatial dimension) in the analysis.

Thus, the GWPCA, considered originally as different localized PCAs computed in spatial locations [9], can be indexed in time in order to assess the changes in the definition of the principal components over time, as formalize below.

Let $\mathbf{A}_t = [A_p(\mathbf{s}_i, t)]$, $i = 1, 2, ..., N$, $p = 1, 2, ..., K$, be the $(N \times K)$ data matrix for the time t, with $K > 2$ variables measured at N spatial locations and $(s_1, s_2, ..., s_d)_i$ are spatial coordinates $(d = 2)$ of the i-th location $\mathbf{s}_i$. The geographically weighted variance-covariance matrix in GWPCA $\mathbf{\Sigma}_t(\mathbf{s}_i)$ is defined as,

$$\mathbf{\Sigma}_t(\mathbf{s}_i) = \mathbf{A}_t^T \mathbf{W}_t(\mathbf{s}_i) \mathbf{A}_t \tag{1}$$

where, $\mathbf{W}_t(\mathbf{s}_i)$ is a (N × N) diagonal matrix of spatial weights for each spatial location $\mathbf{s}_i$, which are generated through a specific user-chosen kernel function, namely a distance-decay weighting function which depends on a properly selected bandwidth parameter. The diagonal entries $\omega_{tj}(\mathbf{s}_i)$ of the weights matrix $\mathbf{W}_t(\mathbf{s}_i)$ correspond to the most used kernel functions [7]. In GWPCA, the matrix $\mathbf{\Sigma}_t(\mathbf{s}_i)$ is decomposed as follows

$$\mathbf{\Sigma}_t(\mathbf{s}_i) = \mathbf{Q}_t(\mathbf{s}_i) \mathbf{\Psi}_t(\mathbf{s}_i) \mathbf{Q}_t(\mathbf{s}_i)^T \tag{2}$$

where $\mathbf{\Psi}_t(\mathbf{s}_i)$ is the diagonal matrix of the local eigenvalues and $\mathbf{Q}_t(\mathbf{s}_i)$ is the matrix of the local eigenvectors, for each time t. In this way, for a fixed time t, there are K eigenvalues and K sets of components' loadings at each spatial points of the area under study.

It is worth to underline that, in this study, the corresponding bandwidth has been fixed on the basis of the structural spatial analysis computed on each variable [10].

3 Study Area

The study area refers to the Emilia-Romagna Region, located in the North of Italy at an altitude of 211 m above sea level. Its area covers *22,510* km^2 from the Appennini mountains to the Po river. Tributaries of the Po river in the western part of the Region (Trebbia, Nure, Taro, Parma, Enza, Secchia, Panaro), are the most abundant of water. The Reno river is the main river of Emilia-Romagna and has several tributaries called Sillaro, Santerno and Senio; other streams such as Fiumi Uniti, Savio, Marecchia and Conca flows directly in the sea. Water is present in different forms among the territory, a complex system of surface and underground sources shapes and characterizes the morphology and the landscape of Emilia-Romagna Region. For this particular territorial structure, the main purpose of environmental management is to ensure the conservation and safeguarding of water throughout the territory of Emilia-Romagna, through strategies for saving and for sustainable water consumption.

Fig. 1. The study area and sample points

The regional government also plans the management of water policies through monitoring and control of surface and groundwater, by observing different physical and chemical parameters. In particular, the data set under study consists of monthly observations regarding macro descriptors of water quality, such as pH, temperature (°C), flow (m^3/s), suspended solids (mg/L), conductivity ($\mu S/cm(20°)$), hardness (mg/L of $CaCO_3$), total nitrogen (mg/L), ammonia nitrogen (mg/L), nitric nitrogen (mg/L), dissolved oxygen (mg/L), BOD5 (mg/L), COD (mg/L), orthophosphate (mg/L), total phosphorus (mg/L), chloride (mg/L), sulfate (mg/L) and Escherichia coli ($UFC/100mL$). The observations are measured at 53 stations of the study area (Fig. 1) and refer to the period from 2010 to 2019 [11].

4 GWPC Analysis

The time indexed GWPCA has been applied to the yearly mean data of the observed variables over the study area. Since the variables have not the same magnitude and some of them have also different units of measure, the data have been standardized. Before applying GWPCA for the each available year, the exponential kernel function has been chosen and the corresponding bandwidth has been fixed on the basis of the structural spatial analysis. Table 1 summaries the results of the ten GWPCA indexed in time and it is evident that the two first local Principal Components (PCs) account together a proportion of the total variance of the data which ranges, in median, from 74.05% to 80.72%. These values are larger than the cumulative portions of variance explained by the first two global PCs (56–63% for the different years) obtained through the standard PCA.

Table 1. Local cumulative portion of variance %

	Minimum	1st quartile	Median	3rd quartile	Maximum
2010	53.15	65.86	74.83	82.25	89.60
2011	54.77	66.58	74.37	77.35	87.09
2012	55.64	67.95	74.05	78.52	87.05
2013	59.20	66.85	75.15	81.85	88.16
2014	50.70	65.52	80.53	82.54	89.74
2015	51.96	63.87	73.34	78.56	87.12
2016	53.26	70.26	74.68	77.48	86.74
2017	53.88	63.49	74.57	78.53	86.96
2018	60.44	69.85	80.72	85.33	92.57
2019	58.93	73.91	80.38	86.17	93.68

By considering the results related to the first local PC, the scores computed can represent a measurement of a composite water quality indicator, as a linear combination of the standardized variables. In order to understand GWPCA's output, it is very useful to identify the winning variable, i.e. the variable with the highest local relative loadings. From 2010, nitric nitrogen and total phosphorus have been almost prevalent in the North of the Region, while the conductibility, ammonia nitrogen and chloride played a main role in the composition of the first PC in the South part. In the last years, other variables have become relevant in the Northern part, such as Escherichia coli and BOD5. In the central part of the Region, different variables have become predominant over time; the suspended solids and the hardness have been important during the first years, then also the Escherichia coli, conductibility and dissolved oxigen have become significant. It is worth pointing out that the loadings concerning the dissolved oxygen are always the opposite of the loadings related to the other variables under study. This is because the dissolved oxygen represents the suitability of water to accommodate

living beings, it has a positive correlation with respect to the water quality; on the other hand, the other variables under study are a measurement of the water pollution and are in contrast with dissolved oxygen.

In the present contribution, the applied GWPCA takes into account the spatial and the temporal dimensions, since it has been indexed in time; however, future studies could propose a comparative analysis with respect to multivariate techniques able to consider simultaneously the spatial and temporal dimensions.

References

1. Brown, D.: Oregon water quality index: background, analysis and usage. State of Oregon Department of Environmental Quality, Laboratory and Environmental Assessment Program (2019)
2. Canadian Council of Ministers of the Environment 2001 Canadian water quality guidelines for the protection of aquatic life: CCME Water Quality Index 1.0, Technical report. Canadian environmental quality guidelines, Canadian Council of Ministers of the Environment, Winnipeg (1999)
3. De Iaco, S., Myers, D., Posa, D.: Total air pollution and space-time modeling. In: Monestiez, P., Allard, D., Froidevaux, R. (eds.) geoENV III, Geostatistics for Environmental Applications, pp. 45–56. Kluwer, Dordrecht (2001)
4. De Iaco, S., Myers, E.D., Posa, D.: Space-time variograms and a functional form for total air pollution measurements. Comput. Stat. Data Anal. **41**(2), 311–328 (2002)
5. Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S.: Principal component analysis on spatial data: an overview. Ann. Assoc. Am. Geograph. **103**(1), 106-128 (2013). https://doi.org/10.1080/00045608.2012.689236
6. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically Weighted Regression - The Analysis of Spatially Varying Relationships. Wiley, Chichester (2013)
7. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P.: GWModel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. **63**(17), 1–50 (2015). https://doi.org/10.18637/jss.v063.i17
8. Hamlat, A., Guidoum, A., Koulala, I.: Status and trends of water quality in the Tafna catchment: a comparative study using water quality indices. J. Water Reuse Desal. **7**, 228–245 (2017). https://doi.org/10.2166/wrd.2016.155
9. Harris, P., Brunsdon, C., Charlton, M.: Geographically weighted principal components analysis. Int. J. Geogr. Inf. Sci. **25**(10), 1717–1736 (2011). https://doi.org/10.1080/13658816.2011.554838
10. Palma, M., De Iaco, S., Cappello, C., Distefano, V.: Tourism composite spatial indicators through variography and geographically weighted principal components analysis. Ann. Oper. Res. 1–19 (2023)
11. Regional Agency for the Prevention and Environmental Protection of Emilia Romagna - ARPAE. https://www.arpae.it/it. Accessed March 2024
12. Samadi, M.T., Sadeghi, S., Rahmani, A., Saghi, M.H.: Survey of water quality in Moradbeik river basis on WQI index by GIS. Environ. Eng. Manag. J. **2**, 7–11 (2015)

Space-Time Analysis of Particle Pollution and Its Effect on Biodiversity

Giuseppina Giungato and Sabrina Maggio(✉)

Department of Economic Sciences, University of Salento, Lecce, Italy
{giuseppina.giungato,sabrina.maggio}@unisalento.it

Abstract. Studying and forecasting spatio-temporal evolution of fine particulate matter represents an essential task because it constitutes one of the main risks to human health and to biodiversity. This work aims to propose an integrated approach based on the combination of a multilevel and geostatistical analysis in order to evaluate the net effect of the atmospheric suspended particulate matter concentrations on the biodiversity occurring in Apulia Region (South of Italy) in the last years, considering the influence of meteorological conditions. The empirical results will demonstrate a relationship between biodiversity, particulate matter concentrations and deterioration of air quality. Moreover, the construction of probability maps will allow to identify low or high biodiversity areas within the domain under study. These evidences can support policy makers in planning strategies aimed at protecting environment and public health.

Keywords: biodiversity · BMI · spatio-temporal analysis · probability maps

1 Introduction

Apulia is a region of the South of Italy which is among the most abundant in the Mediterranean area in terms of biodiversity, due mainly to a favourable geographical position and a wide variety of geological, climate and vegetation conditions. The study of biodiversity is a young science because only in recent times the value of biological diversity has received attention, particularly because of the recognised importance of human being wellness and the maintenance of ecosystem services upon which humanity depends. Nevertheless, this rich biodiversity is under serious threat and a complete survey of biodiversity and an analysis of its spatial patterns from the field shall be associated with satellite imagery, which can provide a synoptic observation in space and time of the territory [2].

In an era of fast and dramatic changes in ecosystems, among the most dangerous pollutants, fine particulate matter represents one of the main threats to human health and to biodiversity (the core of an ecosystem), as it can be released in the atmosphere with concentrations which can dynamically fluctuate across

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 253–258, 2025.
https://doi.org/10.1007/978-3-031-64346-0_43

time and space, owing to the effect of meteorological conditions, physiological features of vegetation, and anthropic activities. As a consequence, studying and forecasting its spatio-temporal evolution is a fundamental task, which can play a key role in the application of the Zero Pollution Action Plan, set out in the year 2021 by the European Commission.

This paper aims to propose a hybrid approach based on the combination of a multilevel and geostatistical analysis in order to evaluate the net effect of the atmospheric suspended particulate matter concentrations on the biodiversity occurring in Apulia Region in the last years, taking into account the influence of meteorological conditions in the area under study. More specifically, the present work has been focused on a) the computation of the Biodiversity Multilevel Index (BMI) [2] of the Apulia Region (South of Italy), b) the spatio-temporal analysis [3,7] of potential environmental factors (i.e. PM_{10} concentrations and climatic variables, such as Temperature, Atmospheric Pressure and Precipitation) which could alter the biodiversity, c) the construction of probability maps that the BMI does not exceed different thresholds, by non-parametric geostatistical techniques, d) the multiple geographical regression analysis of BMI.

2 Materials and Methods

In this section, after a description of the data used for the analysis, the definition of the BMI, obtained on the basis of the species diversity and the landscape characteristics in terms of ecosystems diversity (land cover classes), has been introduced. Moreover, the construction of the probability maps that the BMI does not exceed different thresholds has been discussed, by applying the spatial indicator kriging over the study area [4,9]. In the last part of the section, after the spatio-temporal analysis of PM_{10} and climatic variables, the multiple geographical regression of the BMI has been carried out, by considering the covariates PM_{10}, Temperature, Atmospheric Pressure and Precipitation.

2.1 Data Description

The data used in this study are collected by free public databases [1,5,6] regarding Apulia (South-Eastern of Italy), which is among the richest regions in the Mediterranean area in terms of biodiversity. More specifically, the animal and vegetal species data, available from [5], concerns a total of 729 species (out of which 519 animalae, 4 fungi, 206 plantae) detected in 1,014 locations. On the other hand, PM_{10} (μ/m^3), Temperature ($^{\circ}C$), Atmospheric Pressure ($mmHg$) and Precipitation (mm) are downloaded from [1] and the land use [6] is available from the National Geoportal services, made available by the Ministry of Environment.

2.2 The Biodiversity Multilevel Index

After defining a regular grid of 15×13, with cells size $19.5\,\mathrm{km} \times 19.5\,\mathrm{km}$, the map of the number of species [5] has been overlapped on the land cover map [6]

of the study area in order to identify how they are located in comparison to the changes in land use. Then, the values for the variables S which represents the "true" species richness (effective number of species) and LC which is the "true" land cover richness (effective number of land cover classes), have been computed.

On the basis of these values, the BMI has been obtained through the following ratio:

$$(S \times LC)/A,$$

where A is the sampled area.

The value of the BMI varies from 0 to $+\infty$. The BMI considers the effective species and land cover richness as multiplicative factors that increase the biodiversity of an area. Note that the normalisation against the sampled area allows an easy comparison of the index among different sites.

2.3 Construction of the Probability Maps of the BMI

Given the spatial indicator random field I which takes value 1 if BMI is less equal to 1 for three fixed thresholds z_h, $h = 1, 2, 3$, and equal to 0, otherwise, i.e.

$$I(\mathbf{s}; z_h) = \begin{cases} 1 & \text{if } BMI(\mathbf{s}) \leq z_h, \quad h = 1, 2, 3, \\ 0 & \text{otherwise}, \quad \mathbf{s} \in D, \end{cases} \tag{1}$$

with $z_1 = 1.401$ (Mean), $z_2 = 0.565$ (Median) and $z_3 = 1.598$ (3^{th} Quartile), the indicator kriging allows the estimation of the probability that BMI does not exceed specific threshold values over the domain D under study.

Figure 1 shows the probability maps of the BMI by fixing the three above mentioned thresholds, which highlight that the high probability (values close to 1, with colormap in light grey color) that BMI does not exceed the threshold is prominent in the hinterland, where biodiversity is not guaranteed. On the other hand, the low probability (values close to 0, with colormap in dark grey color) that BMI does not exceed the threshold is evident in the coastal zone, where the richness of species is satisfied (high biodiversity).

2.4 Spatio-temporal Analysis of PM_{10} and Climatic Variables

Based on the data collected within the domain under study, the spatio-temporal analysis of PM_{10} and climatic variables have been conducted by considering the following steps for each variable: a) structural analysis, b) interpolation over a regular grid (the same fixed for BMI), c) average in time in order to obtain an yearly index.

First of all, the structural analysis for each variable has been executed, by estimating and modeling the spatial and temporal marginal variograms through exponential models fitted to each of them [7]. Then, the generalized product-sum models [3] have been fitted to the empirical surface of the spatio-temporal variogram for PM_{10}, Temperature, Precipitation and Atmospheric Pressure, as

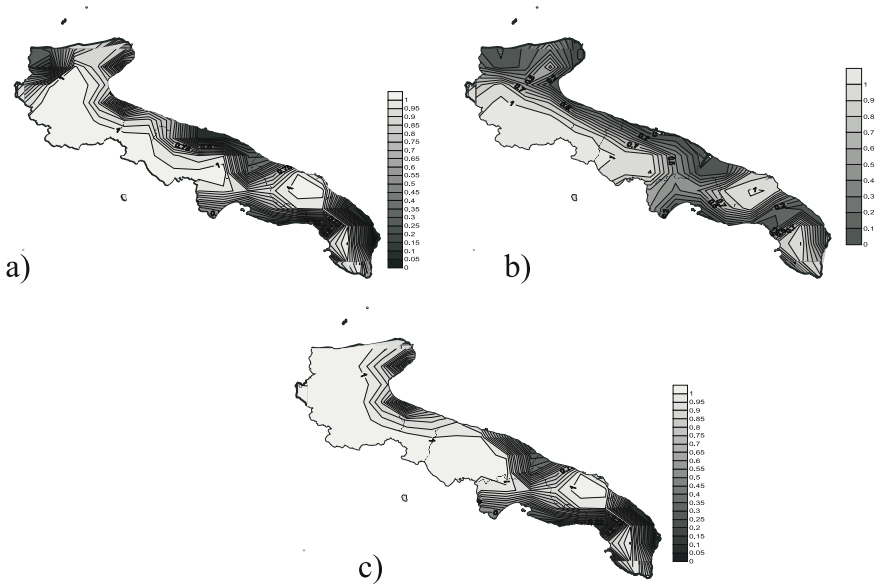


Fig. 1. Probability maps that BMI does not exceed the threshold values: a) Mean; b) Median; c) 3^{th} quartile

shown in Fig. 2. Furthermore, an interpolation of each variable over the previous regular grid (with 19.5 km × 19.5 km cells size), covering the study area, has been performed. Finally, an average in time has been computed for each variable, in order to obtain an yearly index.

2.5 Multiple Geographical Regression of the BMI

In order to analyze the effects of the PM_{10} concentrations and climatic variables, the multiple geographical linear regression for BMI has been performed and the following model has been fitted:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

where the response variable Y represents the BMI and $p = 4$ denotes the number of predictors (covariates) which could alter the biodiversity. In particular, the covariates selected for modeling purposes are the daily average measurements of Atmospheric Temperature, Atmospheric Pressure, Precipitation and PM_{10} concentrations.

From the results reported in the Table 1, it is evident that the covariates selected are all statistically significant. Indeed, the overall goodness-of-fit of the multiple linear regression model is confirmed by the statistics results, expressed in terms of the R-squared (with a value of 0.718), the Multiple R (which takes value 0.744) and Adjusted R squared (with a value of 0.654) indices.

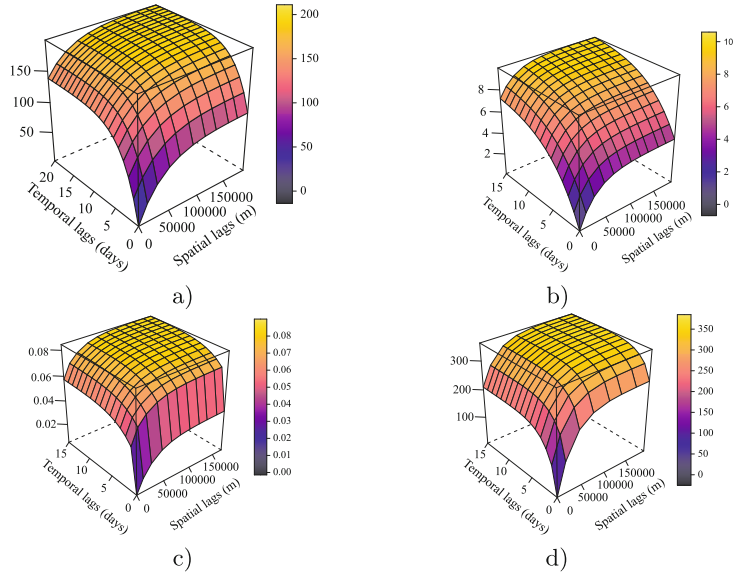


Fig. 2. Fitted product-sum model for: a) PM_{10}; b) Temperature; c) Precipitation; d) Atmospheric Pressure

Table 1. Regression Statistics of multiple linear regression for BMI, together with the estimates of the parameters, the standard errors (SE), the T statistic, the *p-value* and the corresponding confidence intervals (IC) at 95%

Regression Statistics						
Multiple R	0.744					
R squared	0.718					
Adjusted R squared	0.654					
Standard Error of the regression	1.812					
Number of observations	60					
Parameters	$\hat{\beta}$	$SE(\hat{\beta})$	T statistics	*p-value*	[95% IC] Lower	Upper
Intercept	−49.022	16.563	−2.960	0.004	−82.148	−15.896
PM_{10}	−0.109	0.046	−2.370	0.021	−0.201	−0.017
Precipitation	46.118	18.721	2.463	0.017	8.676	83.56
Atmospheric Pressure	0.036	0.017	2.118	0.038	0.002	0.07
Atmospheric Temperature	0.784	0.292	2.685	0.009	0.2	1.368

3 Concluding Remarks

The empirical results demonstrate a relationship between biodiversity (measured by means of BMI), particulate matter concentrations (a key driver of biodiversity decline) and deterioration of air quality. Moreover, the construction of probability maps allow to identify low or high biodiversity areas within the domain under study. These evidences can support policy makers in planning strategies aimed at protecting environment and public health in compliance with the United Nations Sustainable Development Goals of the 2030 Agenda [8].

References

1. ARPA Puglia - Regional Agency of Environment Protection. https://www.arpa.puglia.it
2. Cazzolla Gatti, R., Notarnicola, C.: A novel Multilevel Biodiversity Index (MBI) for combined field and satellite imagery surveys. Glob. Ecol. Conserv. **13**, e00361 (2018)
3. De Iaco, S., Myers, D.E., Posa, D.: Space-time analysis using a general product-sum model. Stat. Probab. Lett. **52**(1), 21–28 (2001)
4. Diodato, N., Bellocchi, G.: Spatial probability modelling of forest productivity indicator in Italy. Ecol. Ind. **108** (2020). https://doi.org/10.1016/j.ecolind.2019.105721.hal-02320568
5. Global Biodiversity Information Facility, GBIF, Free and open access to biodiversity data (2023). https://www.gbif.org/
6. National geoportal services. http://www.pcn.minambiente.it/mattm/servizio-wms/
7. Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University Press, New York (1997)
8. OECD How Was Life? Volume II: New Perspectives on Well-Being and Global Inequality Since 1820. OECD Publishing, Paris (2021). https://doi.org/10.1787/3d96efc5-en
9. Posa, D.: The indicator formalism in spatial statistics. J. Appl. Stat. **19**(1), 83–101 (1992)

Exploring Land Use and Land Cover Changes in Apulia, Italy: Random Forest Approach Utilizing Remote Sensing Data

Iman Masoumi[1,2(✉)], Sandra De Iaco[1,2,3], and Sabrina Maggio[1]

[1] Department of Economic Sciences, University of Salento, Lecce, Italy
{iman.masoumi,sandra.deiaco,sabrina.maggio}@unisalento.it
[2] National Biodiversity Future Center, Palermo, Italy
[3] National Centre for HPC, Big Data and Quantum Computing, Bologna, Italy

Abstract. This study investigates the change of land use and land cover and their impacts on land surface temperature in the Apulia region (South of Italy). Utilizing remote sensing technologies and the Random Forest method, land use and land cover classification in 2013, 2017 and 2023 are analyzed, revealing a significant increase in urbanization and a decline in vegetation cover. The analysis of land surface temperature data has indicated temperature escalations, particularly in urbanized areas, emphasizing the influence of land use and land cover changes on local climate dynamics. The findings underscore the importance of integrating spatial analysis with ecological assessments for informed decision-making in sustainable land management and urban planning. High accuracy in land use and land cover classification validates the reliability of the methodology employed. This study contributes valuable insights into understanding the relationship between human activities, land use and land cover changes, and their environmental consequences, providing a foundation for further research on biodiversity changes and effective mitigation strategies.

Keywords: land use and land cover (LULC) · land surface temperature (LST) · Apulia · Random Forest (RF)

1 Introduction

Sustainable management of natural resources is crucial, given the rapid population growth and shifting climate patterns globally [1]. Human activities like mining, deforestation, agriculture, and urbanization significantly contribute to change Land Use and Land Cover (LULC), impacting ecosystems, biodiversity, water quality and air pollution. Urbanization and population density worsen mentioned issues and lead to higher air temperatures and the creation of Urban Heat Islands (UHI) [2, 3].

Remote sensing technologies, notably Google Earth Engine (GEE), are pivotal for monitoring LULC changes [4]. Landsat 8 Satellite imagery, offering multispectral data, is utilized for LULC classification employing indices like Normalized Difference Built-up Index (NDBI), Normalized Difference Bareness Index (NDBaI), and Normalized

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 259–264, 2025.
https://doi.org/10.1007/978-3-031-64346-0_44

Difference Vegetation Index (NDVI) [2, 5]. Additionally, MODIS data aids in extracting Land Surface Temperature (LST) for analyzing temperature fluctuations [6].

In the recent years, the adoption of Machine Learning (ML) methods for LULC classification has surged due to their efficiency and accuracy [7, 8]. Several studies have investigated the correlation between LULC indices and LST through various methodologies. According to [2, 9], the transformation of vegetated surfaces into impervious areas contributed to the rise in LST. [10] observed that built-up areas foster urban temperature escalation, while areas with abundant vegetation and water bodies exhibit lower LST. [11] highlighted a significant relationship between NDBI and LST, coupled with an inverse correlation between NDVI and LST.

Despite numerous studies scrutinizing the nexus between LULC changes, UHI, and LST across diverse urban landscapes, none of them have explored these phenomena in the Apulia region of Italy. Therefore, the main goals of this study concern 1) the use of remote sensing data and the implementation of RF method for mapping LULC, 2) the analysis of LST based on MODIS data, 3) the investigation of the relationship between LULC changes and LST in the Apulia region for the years 2013, 2017 and 2023. Through the exploration in the Apulia region, this study seeks to augment the knowledge base pertaining to sustainable land management practices and inform decision-making processes concerning environmental conservation and management. Furthermore, this research serves as the initial phase of a broader investigation, with subsequent stages aiming to explore the effect of these changes on biodiversity in the Apulia region.

2 Study Area

The Apulia region, situated in the Southeastern part of Italy, encompasses a diverse landscape characterized by its semi-arid Mediterranean climate, warm summers, and mild winters. Moreover, thanks to its extended area of approximately 19.5 km^2 and more than 4 million inhabitants, Apulia boasts a rich agricultural heritage, ranking second in Italy for the production of olive oil, wine, oats and vegetables. These factors, alongside its geographic features and human activities, intricately influence surface temperatures, presenting a compelling area for comprehensive studies on LULC and LST crucial for sustainable land management and environmental conservation efforts [12].

3 Materials and Method

This study leverages GEE, a robust platform consolidating various Earth datasets including those from MODIS, Landsat 8, and Sentinel, as well as diverse geospatial data encompassing demographics and climate information. GEE provides access to these free satellite images through the United States Geological Survey (USGS) [4, 6].

3.1 Landsat 8 Satellite Image and MODIS Data

The classification of LULC in Apulia during the summer season has utilized Landsat 8 images from 2013, 2017, and 2023. Only images with less than 3% cloud cover have been

selected, and atmospheric correction and mosaicking have been performed. LST data have been obtained from the MYD11A2 V6 dataset in GEE. LST have been derived from MODIS radiance measurements at Bands 31 and 32 from the Terra satellite, providing both day and night LST with a spatial resolution of 1 km [13].

3.2 RF Method

In this study, RF has been applied for LULC classification using Landsat satellite imagery. RF, an ensemble learning algorithm, combines multiple decision trees to enhance classification accuracy. Unlike parametric classifiers, RF doesn't assume a normal distribution of data, making it well-suited for diverse and skewed distributions found in LULC datasets [14]. RF stands out in LULC classification due to its capacity to manage high-dimensional data with intricate variable relationships. It addresses overfitting by selecting random subsets of training samples and features for each decision tree, enhancing generalization [7, 8].

4 Results and Discussion

In this section, the LULC classification and LST maps have been obtained for the Apulia region, and their relationship has been investigated.

4.1 LULC Classification

The NDVI criterion detects vegetation, NDBaI quantifies bare land, and NDBI offers insights into urban climate and ecology [2]. These indices are crucial for mapping LULC based on Landsat 8 spectral bands, as shown in Table 1.

Table 1. Indices for classification LULC [2]

Indices		
$\text{NDVI} = \frac{NIR-RED}{NIR+RED}$	$\text{NDBaI} = \frac{SWIR-TIR}{SWIR+TIR}$	$\text{NDBI} = \frac{SWIR-NIR}{SWIR+NIR}$

This classification by using RF method categorized the LULC into four primary classes: vegetation, built-up areas, bare land, and water bodies. A non-uniform stratified sampling approach was employed to select sample points in Apulia. For these three years analysed, numerous samples have been employed to establish correlations among categorized image cells and reference points. Approximately 200 ground points have been collected for each categorized image, facilitating a comprehensive examination of the precision of LULC classifications.

LULC classification maps are depicted in Fig. 1. Significant changes occurred between 2013 and 2023 in built-up areas, vegetation cover, and bare land. Built-up regions expanded notably, rising from 3.98% in 2013 to 9.90% in 2017 and 19.85% in 2023, while vegetation coverage declined from 37.36% to 31.40% during the same

period. Bare land decreased from 57.61% to 47.64% by 2023, while the water body class reduced from 1.25% in 2013 to 1.11% in 2023 (Fig. 2). This trend signifies increased urbanization and reduced vegetation and bare land attributed to urban expansion due to population growth, migration, and infrastructure development. The evaluation showed overall accuracy 92%, 94.28% and 96.15% for 2013, 2017 and 2023, respectively. Kappa coefficients have been calculated 0.91, 0.89 and 0.92 in 2013, 2017 and 2023, indicating strong agreement between referenced and classified maps.

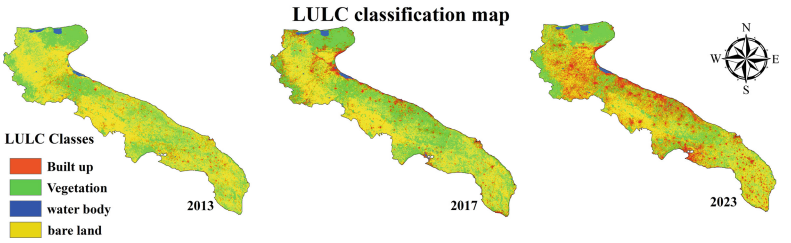
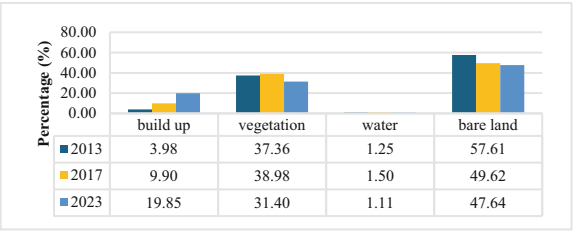


Fig. 1. LULC classification maps in 2013, 2017 and 2023.



	build up	vegetation	water	bare land
2013	3.98	37.36	1.25	57.61
2017	9.90	38.98	1.50	49.62
2023	19.85	31.40	1.11	47.64

Fig. 2. Percentage of LULC classes in 2017 and 2023

4.2 LST

The analysis, utilizing MODIS data from 2013 to 2017 and 2023, has revealed notable changes in LST across Apulia. This is illustrated in Fig. 3, where a gradient from blue to red signifies a transition from lower to higher temperatures. Notably, central and northern parts consistently register higher temperatures during 2023. Urban expansion and vegetation decline likely contributed to temperature increases, particularly evident in the northern regions where temperatures rose by over 4°K between 2017 and 2023 (Fig. 4). The results emphasize the role of LULC changes in exacerbating UHI effects and subsequent changes in LST. Integrating these findings into urban planning and management policies is essential for reducing adverse impacts of urbanization on local climates and enhancing urban livability.

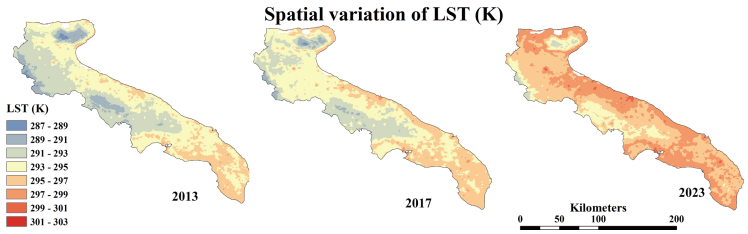


Fig. 3. Spatial variation of LST (˚K) in Apulia for the years 2013, 2017 and 2023

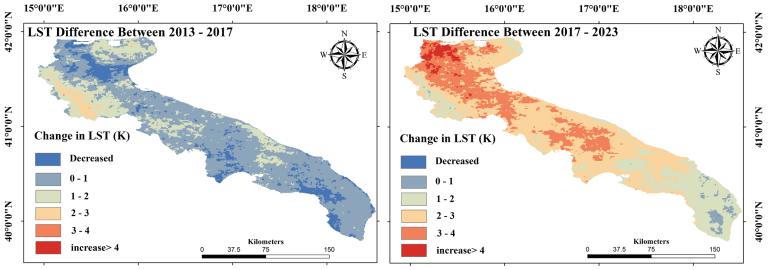


Fig. 4. Changes in LST (˚K) between 2013, 2017 and 2023

5 Conclusions

In conclusion, this study provides valuable insights into the changes of LULC and their implications on LST over the Apulia region (South of Italy). Through the utilization of remote sensing technologies and RF method, significant transformations in LULC were observed over the studied period, notably marked by increased urbanization and declining vegetation cover. The analysis of LST revealed temperature escalations, particularly in urbanized areas, underscoring the influence of LULC changes on local climate dynamics. These findings emphasize the importance of integrating spatial analysis with ecological assessments to inform sustainable land management practices and urban planning policies. Moving forward, further research will be used to explore the cascading effects of these changes on biodiversity and to develop effective mitigation strategies for mitigating adverse environmental impacts.

Acknowledgements. This research was supported by the National Biodiversity Future Center-NBFC, Spoke 4, Activity 4.1, sub activity 4.1.1, in the research Programme "*Analisi geostatistica delle risposte ecologiche agli impatti*".

Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – Next GenerationEU.

Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP F87G22000290001, Project title "National Biodiversity Future Center - NBFC".

References

1. Hunt, J.C.R., Aktas, Y.D., Mahalov, A., Moustaoui, M., Salamanca, F., Georgescu, M.: Climate change and growing megacities: hazards and vulnerability. Proc. Inst. Civ. Eng. Eng. Sustain. **171**(6), 314–326 (2018)
2. Hamed Fahmy, A., Amin Abdelfatah, M., El-Fiky, G.: Investigating land use land cover changes and their effects on land surface temperature and urban heat islands in Sharqiyah Governorate, Egypt. Egypt. J. Remote Sens. Space Sci. **26**(2), 293–306 (2023)
3. Tan, J., et al.: The urban heat island and its impact on heat waves and human health in Shanghai. Int. J. Biometeorol. **54**(1), 75–84 (2010)
4. Pande, C.B.: Land use/land cover and change detection mapping in Rahuri watershed area (MS), India using the Google earth engine and machine learning approach. Geocarto Int. **37**(26), 13860–13880 (2022)
5. Elbeih, S.F., El-Zeiny, A.M.: Qualitative assessment of groundwater quality based on land use spectral retrieved indices: case study Sohag Governorate, Egypt. Remote Sens Appl. **10**, 82–92 (2018)
6. Xie, S., Liu, L., Zhang, X., Yang, J., Chen, X., Gao, Y.: Automatic land-cover mapping using landsat time-series data based on Google earth engine. Remote Sens. **11**(24), 3023 (2019)
7. Arpitha, M., Ahmed, S.A., Harishnaika, N.: Land use and land cover classification using machine learning algorithms in google earth engine. Earth Sci. Inform. **16**(4), 3057–3073 (2023)
8. Traoré, F., Palé, S., Zaré, A., Traoré, M.K., Ouédraogo, B., Bonkoungou, J.: A comparative analysis of random forest and support vector machines for classifying irrigated cropping areas in the Upper-Comoé Basin, Burkina Faso. Indian J. Sci. Technol. **17**(8), 713–722 (2024)
9. Pal, S., Ziaul, S.K.: Detection of land use and land cover change and land surface temperature in English Bazar urban centre. Egypt. J. Remote Sens. Space Sci. **20**(1), 125–145 (2017)
10. Ogashawara, I., Bastos, V.: A quantitative approach for analyzing the relationship between urban heat islands and land cover. Remote Sens. **4**(11), 3596–3618 (2012)
11. Chen, X.L., Zhao, H.M., Li, P.X., Yin, Z.Y.: Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sens. Environ. **104**(2), 133–146 (2006)
12. Serio, F., et al.: Groundwater nitrate contamination and agricultural land use: a grey water footprint perspective in Southern Apulia Region (Italy). Sci. Total. Environ. **645**, 1425–1431 (2018)
13. Wan, Z., Hook, S., Hulley, G.: MYD11A2 MODIS/Aqua Land Surface Temperature/Emissivity 8-Day L3 Global 1 km SIN Grid V006. Distributed by NASA EOSDIS Land Processes DAAC (2015). https://doi.org/10.5067/MODIS/MYD11A2.006. Accessed 23 June 2019
14. Traoré, F., Cornet, Y., Denis, A., Wellens, J., Tychon, B.: Monitoring the evolution of irrigated areas with Landsat images using backward and forward change detection analysis in the Kou watershed, Burkina Faso. Geocarto Int. **28**, 733–752 (2013). https://api.semanticscholar.org/CorpusID:129540312

The Narrative of Gender and the Profound Impact of Language Weight

Maria Giuseppina Muratore(✉), Claudia Villante, and Lucilla Scarnicchia

Italian National Institute of Statistics, Directorate for Social Statistics and Population Census, Rome, Italy
{muratore,Claudia.villante,Lucilla.scarnicchia}@istat.it

Abstract. The stereotypical nature of society serves as the foundation for gender discrimination against women, hindering the attainment of equal opportunities and impeding the complete emancipation of women from men.

Verbal and written communication plays a fundamental role in perpetuating gender stereotypes. The words and how they are used can influence the perception and treatment of women in society. The importance of language in the context of gender discrimination emphasizes the need to carefully examine how words contribute to the construction of harmful stereotypes and how they can be consciously used to promote positive change.

Additionally, these stereotypes constitute a primary catalyst for gender-based violence against women. In this context, delving into the study of gender role stereotypes becomes synonymous with investigating the roots of violence, understanding how it persists, and identifying the challenges that impede efforts to eradicate it.

This paper aims to explore the sources employed by Istat to delve into the framework that underlies gender discrimination and other related aspects. Adopting a multi-source approach, the study utilizes data from population surveys and big data to focus on discerning the key determinants of stereotypes. The ultimate goal is to identify strategies to combat these stereotypes and address gender-based violence effectively.

Keywords: Gender stereotypes · attitudes towards violence · social media · language

1 Introduction

Gender stereotypes concern beliefs about what men and women typically do and about what they should do. They are descriptive and prescriptive simultaneously and permeate all aspects of social life conditioning, for instance, educational and occupational choices, career opportunities, the access to the political arena, as well as the places where they decide to live.

The Convention on the Elimination of All Forms of Discrimination against Women adopted by the United Nations General Assembly on December 18, 1979, resulted from over thirty years of work by the UN Commission on the Status of Women, established in

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 265–270, 2025.
https://doi.org/10.1007/978-3-031-64346-0_45

1946 to monitor and promote women's rights. Addressing civil rights, legal status, human reproduction, and cultural factors' impact on gender relations the Convention acknowledges existing discrimination and mandates appropriate measures, including legislation, to ensure women's full development and advancement, ensuring their exercise of human rights on an equal basis with men. The Convention remains highly relevant, as evidenced by the outcomes of the latest periodic review to which Italy was subjected on February 1, 2024. This underscores the ongoing need to monitor and address issues related to gender discrimination, emphasizing the importance of implementing concrete measures to ensure effective equality of rights and opportunities for women. Namely Art.5 clearly refer to the need of eradicating prejudices and practices which are based on the idea of the inferiority of one sex or on stereotyped roles and the need that family education includes a proper understanding of maternity as a social function and the recognition of the common responsibility of men and women in the upbringing and development of their children.

Also the Council of Europe Convention on Preventing and Combating Violence Against Women and Domestic Violence (2011), better known as the Istanbul Convention, recognises that gender stereotypes contribute to making violence against women acceptable in societies and therefore requires promoting a cultural change in their regard. The Convention focuses on stereotyping as a major cause of violence. The gender-based violence (GBV) in fact, is mainly rooted in the culture of disparities and unequal power between men and women that is at the bases of our societies. Even the dualism between the public and private spheres that distinguishes men's and women's lives – representing family life and the home as the safest places for women, as opposed to the street and the city which are seen as risky - is a stereotype and shatter against a very different reality, as demonstrated by the data collected on violence against women.

Istat started to study gender-based stereotypes carrying out in 2018 (the second edition is now ongoing) a dedicated module on gender role stereotypes and the social image of violence, in the context of the agreement with the National Department of Equal Opportunity (Italian Presidency of Council), and, in 2020, developing out an experimental analysis of big data on gender based violence and stereotypes.

Based on these data, firstly, our goal is to measure how gender stereotypes are widespread, looking at differences in the population and trying to identify if there are protective and predictive risk factors for being stereotyped.

Secondly, we wanted to focus on the growing reach of the Internet. The rapid spread of mobile information and communications technologies (ICTs), and the wide diffusion of social media, especially during the pandemic period, have spread online gender-based violence. Consequently, our further questions are: how do these new forms of modern conversation affect gender stereotypes? How gender stereotypes should be studied from this perspective? How to observe what the users think, say and share, and how to monitor the discussions about gender stereotypes on the web?

These relevant and complex questions will be addressed considering different data sources, as described in the following.

Narrative the gender means to look at stereotypes, means to look at discrepancies between men and women, but also at what affects the cycle of gender based violence.

Data results from literature[1] and surveys (Istat, 2008; 2015) underlined the importance of intergenerational transmission of violence. The violent context of the family of origin is associated in fact with the level of violence that characterises the current history of abuses, the violence tolerance as a cause of violence, (Corazziari and Muratore, 2013, pp. 88–116), since women are socialised to consider violence as normal. The language also in this case is crucial. Women consider suffered violence as something wrong but not as a crime and this is one of the reason that create difficulties to get out of it. This message assumes the normality of the men's sexual needs that have to be satisfied, and on the contrary, the same messages assume that women's willing is to be denied. A woman is, by definition, a weak actor in life, an object for the other's wishes that likes to be conquered.

Over time the situation has been changing and the awareness increased. Data support these changes: doubled the women recognised violence suffered as a crime, reported to the police the suffered violence, sought help in anti-violence centres (Istat, 2015), more than doubled the requests for help to 1522, the national helpline against violence against women (by phone and chat) (Istat, 2022), but the recent results from the module on gender stereotypes (2023) confirm a stereotyped society, even if bettered compared to the past.

2 Methods of Analysis

2.1 Surveys Data: The Module on Gender Stereotypes

The module addressing gender role stereotypes and, for the first time in Istat, opinions on the acceptability of violence, its permeation and its causes, as well as stereotypes about sexual violence was carried out in 2018 and in 2023. The questions were addressed to individuals aged 18 to 74 years, interviewed with the CATI (Computer assisted telephone interview) and CAWI (Computer assisted web interview) technique.

The module was organised into six main areas in order to reach specific purposes: stereotypes about gender roles; the acceptability of intimate partner violence; the perception of the extent of the violence; the causes of intimate partner violence; stereotypes about sexual violence; understanding how women and men think of each other, and themselves.

2.2 The Big Data Sources: Machine Learning Method

In 2020 Istat started an experimental study using Big Data (methodology is still ongoing) intended to analyse and monitor the different uses of social media: when the main effect is raising awareness about gender stereotypes and gender-based violence or, on the opposite, when they lead to reinforce the related stereotypes. An additional reason to further develop methods of analysis of social media contents is the fact that they can also

[1] Baldry and Ferraro (2008); Baldry (2007); Corazziari and Barletta (2012); Culross (1999); Cummings, Pepler and Moore (1999); Dauvergne and Johnson (2000); Prinz and Feerick (2003); Baldry, Ferraro and Ferraro (2011).

be used to perpetrate some forms of abusive language, as hate-speech and cyberviolence, like the bodyshaming.

In the experimental study, the contents of social media (Twitter now X, Facebook, Instagram, press review websites) were selected on the basis of specific keywords. They were processed using a machine learning process, which uses a supervised text classification method based on machine learning algorithms. This allowed us to apply the sentiment and emotion analysis to the contents of interest.

The machine learning process has accomplished 3 phases of work: 1) Definition of the requirements of the study and of the criteria for extraction by processing and assembling the corpus of the annotation 2) Setting the dataset 3) Identification of trained classification models to carry out the sentiment and emotional analysis. The classification models implemented are based on an algorithm called Bidirectional Encoder Representations from Transformers (BERT) proposed by Google researchers, part of the Google Brain and Google Research departments.

3 Some Evidence from Results

Results show women are becoming increasingly conscious of gender roles. Noteworthy stereotypes related to gender roles encompass notions like 'men are less suited to do housework than women' (21.4%), 'a woman's fulfillment hinges on having children' (20.9%), 'success in the workplace holds greater importance for men than women' (20.4%), 'mothers bear the primary responsibility for tending to children and their daily needs' (20.2%), and 'it is mainly up to men to provide for the family's financial needs' (17.2%). Less common stereotypes involve beliefs such as 'men wield the utmost influence in making family decisions' (6.3%) and 'a supportive wife/partner should align with her husband/partner's ideas even if she disagrees' (6.5%) (Istat, 2023). In contrast to 2018, all identified gender role stereotypes have witnessed a decline, particularly as perceived by women**.** Nevertheless, data show the persistence of stereotypes on sexual violence and the victims blaming attitudes, the latter a behaviour sometimes perpetrated also by women.

Based on the data collected from social media, by using the machine learning process mentioned above, we can consider quite relevant the volume of conversations related to gender stereotypes, 438.386 buzz, coinciding with related tweets, during an year of observation from May 2022 to May 2023.

Which are the events that generate these conversations? Even if there are some external events that polarize the conversation (such as the March 8 celebration of Women's Day), it is interesting to see how the debate on gender stereotypes achieves a peak around the social role of mother and father stimulated by a TikTok video. The topic addressed regards the traditional female and male roles in a "traditional" family against new forms of families (recorded around the 12th of February 2023). The volume of likes, comments, and retweets has generated an increase in the volume of conversation that the blue line on the chart describes, more linked to this tweet rather than the volume of discussion around the 8th of March woman's celebration day (Fig. 1).

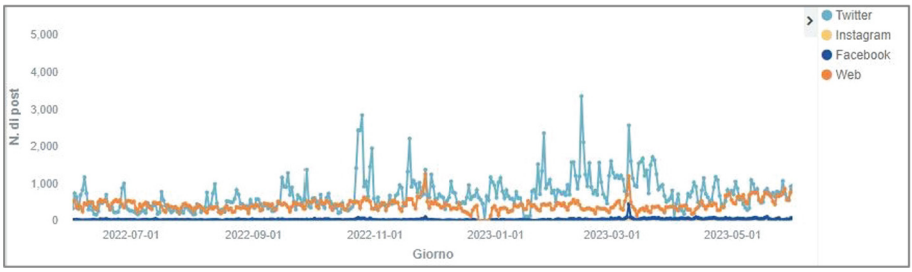


Fig. 1. Number and trend of contents on Gender Stereotypes by social media (31st of May 2022 to 31st of May 2023). Source, Istat 2023 – Sentiment analysis on Gender Stereotypes' posts

4 Conclusion

The paper will produce a multidimensional analysis of data on gender role stereotypes and the social image of violence, with the aim to highlight situations diversified based on the positions about gender roles and sexual violence, the acceptability of violence and the cultural vision people have of gender-based violence.

The objective extends to identifying the determinants of stereotypes to inform the design of more effective policies.

Additionally, we seek to integrate these data with information from the labor market and decision-making realms, crucial aspects for women's emancipation, exploring social classes' differences, urban/rural contexts and territory as strategies of analysis, to add depth to our understanding.

In fact, for example, data underline that the position held at work seems to have an influence on opinions about gender roles, highlighting more stereotyped ideas among labourers and workers with less responsabilities.

Data shown increased awareness among women. However, monitoring gender stereotypes reveal that interpretations of violence and gender roles vary among women due to differences in education, cultural competencies, and generational cohorts. These varying interpretations affect women's ability to emancipate themselves, with some gaining self-awareness and higher self-esteem, while others face unresolved problems.

The analysis of social media discussions is intriguing, although the integration of these data with traditional sources is still evolving. Social media analysis underscores the persistent presence of gender stereotypes, particularly regarding the roles of women and men. Sentiment analysis indicates that social contents generating discussions often express positive sentiments, aligning with traditional views of family roles.

Despite some improvements, implicit barriers to human rights and equal participation in public/economic life still remain entrenched in our society. The study of gender stereotypes in different areas of public life, such as women in politics, education (with particular reference to STEM), the economy, and digital society, is very important, and data on these dimensions are often lacking. Notably, the gender dimension is often absent from discussions on emerging challenges like climate change, urban policy, and the digital society.

References

Baldry, A.C.: "It does affect me" disruptive behaviours in preadolescents directly and indirectly abused at home. Eur. Psychol. **12**(1), 29–35 (2007)

Baldry, A.C., Ferraro, E.: Uomini che uccidono. Storie, moventi e investigazioni. CSE, Torino (2008)

Baldry, A.C., Ferraro, E., Porcaro, C.: Donne uccise e donne maltrattate. Stesso passato ma anche stesso destino? Rassegna Italiana di Criminologia (4), 13–21 (2011)

Corazziari, I., Barletta, R.: The intergenerational transmission of domestic violence: an analysis of data from the Italian "Women Safety Survey." Interdisc. J. Fam. Stud. **XVII**(1), 113–136 (2012)

Cornelli, R.: Pregiudizi, stereotipi e potere. Alle origini delle pratiche di disumanizzazione e delle politiche dell'odio. Rassegna Italiana di Criminologia (3), 206–216 (2019). https://doi.org/10.7347/RIC-032019-p206

Corazziari, I., Muratore, M.G.: Domestic violence: short and long term consequences. La camera Blu, Rivista di studi di genere, Journal of Gender Studies 88–116 (2013). http://www.cameralblu.unina.it/index.php/camerablu/article/view/2821/0

Culross, P.L.: Health care system responses to children exposed to domestic violence. Future Child. Domest. Violence Child. **9**(3), 111–121 (1999)

Cummings, G.J., Pepler, D.J., Moore, T.E.: Behavior problems in children exposed to wife abuse: gender differences. J. Fam. Violence **14**(2), 133–156 (1999)

Dauvergne, M., Johnson, H.: Children witnessing family violence. Juristat Canadian Centre for Justice Statistics, Statistics Canada, no. 85-002-XPE, XXI(6), 1–13 (2000)

Istat: La violenza contro le donne. Indagine multiscopo sulle famiglie "Sicurezza delle donne" anno 2006, Collana Informazioni, Roma (2008)

Istat: Violence against women in and outside the family, year 2014 (2015). https://www.istat.it/it/files//2019/11/violence-against-women-_2014.pdf

Istat: Gender roles, stereotypes and attitudes to sexual violence (2019). https://www.istat.it/en/archivio/236678

Istat: Il sistema di protezione per le donne vittime di violenza - anni 2020 e 2021 (2022). https://www.istat.it/it/archivio/270509

Istat: Stereotipi di genere e immagine sociale della violenza - primi risultati (2023). https://www.istat.it/it/archivio/291163

Muratore, M.G., Scarnicchia, L., Villante, C.: Looking at gender stereotypes to fight gender based violence. Rassegna Italiana di Criminologia **3** (2023). https://doi.org/10.7347/RIC-032023-p211%20

Prinz, R.J., Feerick, M.M.: Next steps in research on children exposed to domestic violence. Clin. Child Fam. Psychol. Rev. **6**(3), 215–219 (2003)

Inefficiencies in Digital Advertising and the Threats of Artificial Intelligence

Giorgio Tassinari(✉)

Department of Statistics "Paolo Fortunati", University of Bologna, Via Belle Arti 41, 40134 Bologna, Italy
giorgio.tassinari2@gmail.com

Abstract. The paper deals with growing importance of digital advertising nowadays and with the inefficiencies that are linked to its measurement and use. It also deals with danger concerning the diffusion of artificial intelligence for consumers independence.

Keywords: Advertising · Artificial Intelligence · Inefficiency

From long time market researchers, firms' executives and managers ask themselves about effectiveness of marketing policies and strategies in competitive environments. Establishing quantitative instruments for evaluating effectiveness is an activity that has received great attention in last years, specially in market of frequently purchased consumer goods, for the availability of scanner data coming from longitudinal samples (panel) of households and points of sale.

In more recent times, internet revolution has made that digital advertising makes available unprecedented innovations for "marketers" (Gordon et al. 2020). Firms now can address their advertising yo very specific segments of individuals, using personalized messages. Comparing digital advertising with traditional one, the first implies a better segmentation, personalized ad contents and the measurement of advertising exposure of the single consumer (Gordon et al., 2020). This caused a strong change in the structure of advertising investments: in Italy in 2020 expenditure for digital advertising overcame that for TV advertising, and it was the first media for investments amount.

For managers' decisions, it important to split short time effects of advertising from medium and long ones. The topic of medium and long term market response to marketing mix activities is at the centre of every marketing strategy that aims to build a competitive advantage for the firm or the brand. As long time series has become available, has rised a strong interest for the analysis of marketing mix impact and for disentangle short and long time effects. The dynamic impact of marketing mix variables on performance's measures as sales or market share has received a great attention in literature.

For giving an appropriate conceptualization of the issues that are linked to measurement of advertising effects, it is possible to advance a simplified scheme that entails three level of consumer response to advertising action:

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 271–274, 2025.
https://doi.org/10.1007/978-3-031-64346-0_46

– message perception and its memorization, that is cognitive response;
– the impact of advertising on the perception of product's attributes, on their evaluation on the formation of preferences for products and brands (affective response);
– the impact on buying and rebuying behaviour (behavioural response).

As for any other form of investment, the measurement of advertising effectiveness is an important issue, also if being only one of the factors that influence purchase behaviour, its contribution cannot be directly estimated. For evaluating advertising effects on sales it is necessary to consider all the influent variables, in a dynamic way if possible.

Anyway, evidence suggests that firm data not always aligns with long term growth goals. (Yuxing Du et al., 2021). The "streetlights" effect is a danger for this convergence (Yuxing Du et al., 2021), effects that origins from the excess of managers' confidence in data readily available because of the easyness of measurement and use. For example the "streetlights effect" caused by the abundance of advertising data made managers giving more attention on advertising management respect to distribution or product line notwithstanding the advertising elasticity is much less respect to the other two (Hanssens and Pauwels, 2016).

We concentrate on web advertising. Internet diffusion in Italy is very wide and internet use is very strong, so digital media is the new horizon of advertising investments. According to Audiweb press release concerning January 2024, web audience (from 2 years old) has been of 44,2 millions of users, the 75,8% of the same age population. The average time of connections in the average day has been of 2 h and 52 min (the average day indicates the number of unique daily users that in a day have made an access to internet. Unique users are the number of single people that have made at least an access to internet). The use of internet by mobile has reached the 81,3% of the population aged 18–74 years.

The methods for measuring web audience are of two types: site-centric, based on the recording of the accesses to the publisher server or to advertising banner and user-centric, based on the recording of user's web activities of a probabilistic sample of internet users through meters installed on pc or mobile. Data obtained by site-centric method allows to know the traffic volume of the site and the analysis of the validity of contents, but don't allow the measurement of the number of people that have visited it. The user centric method allows di know the navigation behaviour of users and to know their characteristics. This method also allows to individuate unique users.

From more than 20 years advertising industry has seen internet as the new communication channel available both for transmit in traditional way both in a interactive way advertising communication. In first instance we can draw a correspondence between the concept of internet advertising with each form of communication that enter with advertising definition and that can be implemented by web communication.

Web advertising is a complex phenomenon, that overcomes the ways in which communication is used on traditional media (push, not explicitly demanded by receiving people), spreading personalized communication. The distinction between communicative-relational process and commercial-transactional process is shadowing. Infact internet is able to expand its communication capability assemblying traditional marketing channels and communication instruments: advertising, public relations, direct marketing, sponsorships. Events. In this contest two issues are very relevant and interdepend: the issue

of control of the communication process between users and advertisers and that of the double modality, push and pull, of internet communication.

First question relates to the capability of internet environment to facilitate the riequilibrium of communication control, making it more fair for the users. On the web, while a part of advertising messages still remains under advertiser's control, a always more consistent part is under user's control, who can chooses a site or another, activate automatic ad blocks, asking informations, contact the advertiser in real time, buy the proposed product. On the contrary, an advertising communication not requested or that interrupts the navigation in a forced manner may be considered intrusive and has null effects or worst.

Clickstream data that track customers across digital channels have afforded marketers a near 360-degree view of a customer's journey online. This has led to tools that support growth by improving marketing return on investment (ROI) through better target selection and media planning, including multitouch attribution models. However, touchpoints are not equally trackable across channels. Discrepancies in measurability between digital and mass media (e.g., TV, radio, print, outdoor) may have contributed to shifts of advertising budgets toward digital due to (1) larger measurement errors in mass media and (2) the difficulty in quantifying the generative influences of mass media and cross-media synergies.

Social media data and other forms of UGC (User Generated Content) have revolutionized the way marketers listen to their customers, dwarfing the data available through traditional tools such as surveys and focus groups. Popular UGC platforms use such data to target advertising. However, the size, timeliness, and richness of UGC does not guarantee that these data are representative. Relying on social listening could bias perceptions of the marketplace due to differences in users across platforms (Schweidel and Moe 2014; Schoenmueller, Netzer, and Stahl 2020).

Many firms have jumped on the big data–machine learning bandwagon as open-source algorithmic methods become readily implementable, without fully grasping the relative advantages of traditional methods taken by marketing science or recognizing the potential for "algorithmic biases" and other unintended consequences (e.g., Lambrecht and Tucker 2019). In pursuing predictive ability, big data applications risk sacrificing the interpretability of the results. Consequently, marketers could inadvertently create social ills in their pursuit of growth that can harm society and, eventually, the firm.

Until few times ago, Artificial Intelligence (AI) was uniquely the object of science fiction. Now it is changing the way in which consumer eat, sleep, work, play and also date. While AI can improve consumes' life in many practical and important issues, notwithstanding it can be a failure in taking into account the behavioural impacts of consumers experience.

In popular culture, the property of personal data is frequently associated with a lose of personal control that takes origins from the dangerous potential of technology that allows the control of human behaviour. Many literately works (from da *1984* by George Orwell) describe systems of oppression in which, because of lack of privacy and of constant surveillance, people are non more able to control their own destinies. This dystopic imagine reflect in scholarly works that associate "data capture" with the rise of capitalistic market in which information became the central form of capital (Zuboff, 2019).

Consumer data are the basis of a new form of business leaded by the ability to colonize private consumer experience. This business contributes to a market of surveillance in which surplus deriving by data in transformed into predictive products (Zuboff, 2019). Under this point of view not only technological firms are continuously searching of new ways for making monitoring and surveillance acceptable by consumers linking them to convenience, productivity, security, health and welfare, but firms have to push limits of the private information that consumers have to share. As consumer behaviour modifies, Ai can transform consumers in subjects that are collaborative with commercial exploitation of their private experience, and in this way lowering personal control and making possible concentration of knowledge and power in the hands of those who own their informations.

References

Du, R.Y., Netzer, O., Mitra, D.: Capturing marketing information to fuel growth. J. Mark. **85**(1), 163–183 (2021)

Gordon, B.R., Jerath, K., Katona, Z., Narayanan, S., Shin, J., Wilbur, K.C.: Inefficiencies in digital advertising markets. J. Mark. **85**(1), 7–25 (2021)

Hanssens, D.M., Pauwels, K.H.: Demonstrating the value of marketing. J. Mark. **80**(6), 173–190 (2016)

Lambrecht, A., Tucker, C.: Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag. Sci. **65**(7), 2966–2981 (2019)

Schoenmueller, V., Netzer, O., Stahl, F.: The polarity of online reviews: prevalence, drivers and implications. J. Mark. Res. **57**(5), 853–877 (2020)

Schweibel, D.A., Moe, W.W.: Listening in on social media: a joint model of sentiment and venue format. J. Mark. Res. **51**(4), 387–401 (2014)

Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books, New York (2019)

Real-Time Anomaly Detection of Spatial Processes via Functional Conformal-Prediction Bands

Teresa Bortolotti, Alessandra Menafoglio, and Simone Vantini(✉)

MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy
{teresa.bortolotti,alessandra.menafoglio,simone.vantini}@polimi.it

Abstract. Temporally variant data observed on two-dimensional domains arise naturally across several disciplines. Functional data analysis proves to be inherently suitable for representing and modeling this kind of data, offering a rigorous mathematical framework capable of preserving the spatially continuous nature of these data. Within this framework, discrete-time evolving surfaces can be effectively modelled as functional time series of random real-valued functions defined on a two-dimensional domain. Building upon this approach, an anomaly detection method for handling such data is here developed. The proposal hinges on a probabilistic forecasting scheme for two-dimensional functional time series that incorporates conformal prediction bands for functional data. This methodology allows real-time construction of a prediction range for each point of the spatial domain ensuring joint control of the coverage probability. An anomaly is identified every time the observed surface deviates from the prediction bounds at a particular point in the domain. This approach inherently guarantees exact control over the probability of encountering one or more false warnings in the spatial domain, offering a viable solution for real-time monitoring of high-resolution spatial data. Finally, the proposed anomaly detection procedure is applied to a dataset collecting weekly interferometric measures of land elevation speed in the Phlegraean Fields volcanic area in Italy.

Keywords: Anomaly detection · functional data analysis · conformal prediction · space-time data

1 Introduction

Functional Data Analysis [8,18] is a field of Statistics focusing on the statistical analysis of data sets made of functions. Originating from pioneering work by Jim O. Ramsay [17], numerous authors have made significant contributions to this field. Works such as Functional Boxplots [20] or Functional Linear Regression [18] exemplify the still lively nature of this area of modern Statistics characterized by many still open research problems. One such challenge is the creation

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 275–280, 2025.
https://doi.org/10.1007/978-3-031-64346-0_47

of prediction sets for newly observed functional data, a task we address in the first part of this work. Our first goal is indeed to develop a method capable of producing valid prediction sets (i.e., guaranteeing a coverage larger or equal than $1 - \alpha$) independently of the sample size and on the unknown distributional model. Within the framework of functional data analysis, this challenge has been mostly approached in two main ways. The first approach relies on parametric bootstrapping techniques e.g. [3,5], while the second method utilizes dimensionality reduction techniques e.g. [2,12]. However, both approaches have limitations: the former relies on specific distributional assumptions and/or asymptotic results while the latter is affected by the approximation introduced by dimensionality reduction. To address these shortcomings, our first contribution focuses on presenting a novel procedure that overcomes these challenges through a new approach in the realm of Conformal Prediction [22].

In the second part of this work, building on the previous approach, an anomaly detection method is presented for handling space data. The central focus revolves around a recent variation of conformal prediction for functional time series [1]. The proposal hinges on a probabilistic forecasting scheme for two-dimensional functional time series, extending a functional autoregressive process of order one to this context and incorporating conformal prediction bands for functional data previously introduced. This methodology allows the real-time construction of a spatially varying prediction bounds for each point of the spatial domain ensuring global control of the coverage probability. The latter represents the probability that the newly observed surface is contained within the prediction range for each point in the domain. An anomaly is then identified every time the observed surface deviates from the prediction range at a particular point in the domain. This approach inherently guarantees exact control over the probability of encountering one or more false warnings across the spatial domain, offering a viable solution for real-time monitoring of high-resolution spatial data. Finally, to illustrate its applicability, the proposed anomaly detection procedure is applied to a dataset that collects weekly interferometric measures of land elevation speed in the volcanic area of Phlegraean Fields (Italy).

A detailed description and a deep theoretical study of our proposal together with further simulations and details about the application can be found in [6,7], and [1]. An introduction to Conformal Prediction can be found instead in the review paper [9].

2 Conformal Prediction

Our contribution will start with a gentle introduction to Conformal Prediction in the univariate setting [19,22]. This approach is an innovative nonparametric approach to create prediction sets which was firstly developed in the field of Machine Learning as a method to construct prediction intervals for Support Vector Machines [11] and already used also in the functional context via the use of a finite-dimension truncated basis expansion [15]. We will show that the core of the approach is the choice of a *nonconformity measure*, namely any measurable

function that takes values in $\mathbb{R}^+$ and whose objective is to score the "extremity" of an observation with respect to the other ones. Specifically, we will restrict our focus to the so called Split Conformal approach which allows the computationally efficient construction of finite-sample valid prediction sets under the assumption of exchangeable data by using a "virtual resampling" kind of reasoning.

3 Conformal Predictions Bands for Functional Data

We will then move to the case of independent and identically distributed functional data, which is our first novel original contribution to the existing literature. We will list the properties that a prediction set is expected to satisfy in the functional framework. We will deepen a very relevant aspect of practical interest which concerns the shape of the sets: in particular, we will show how our functional prediction sets have the geometrical shape of a band. This allows indeed an easy visualization of the prediction set in the natural graphical visualization of functional data [14–16].

After having introduced Split Conformal prediction for univariate data and the properties required to functional prediction set, we will introduce a new group of nonconformity measures for functional data based on the supremum metric allowing the construction of functional prediction bands. These distribution-free functional bands will be proven to be finite-sample valid and defined in closed form. Some emphasis will also be placed on the computational cost characterizing the method. The procedure is indeed highly scalable as the computational effort required by the procedure increases only linearly with the sample size. We will also focus on the width of the prediction bands returned by the procedure. Different nonconformity measures - belonging to the aforementioned group of nonconformity measures - will be compared in different scenarios through simulation studies.

4 Extension to the Real-Time Monitoring of Two-Dimensional Functional Time Series

We will then extend our proposal to the scenario of temporally dependent functional data with possible scalar, categorical, or functional covariates. In detail, we will extend the theory of autoregressive processes in Hilbert spaces in order to allow for real-valued functions with a bivariate domain by integrating this framework [13] with the work [4] which focuses on the construction of conformal prediction intervals in the framework of scalar time series. We will show that this extension is feasible, theoretically easy, and still computationally affordable.

Finally, we will show that in the context of space-time data, this methodology can be fruitfully used to create real-time prediction bounds that vary spatially for each point within the spatial domain, ensuring joint control over the coverage probability. By systematically point-wise comparing at each timestamp the prediction bounds with the newly observed data, anomalies can be identified

whenever the newly observed surface deviates from the prediction range at a specific point in the domain. The joint control of the coverage probability of the prediction range inherently ensures exact control over the probability of detecting one or more false warnings across the spatial domain, regardless of its size. This offers a practical solution for real-time monitoring of high-resolution spatial data.

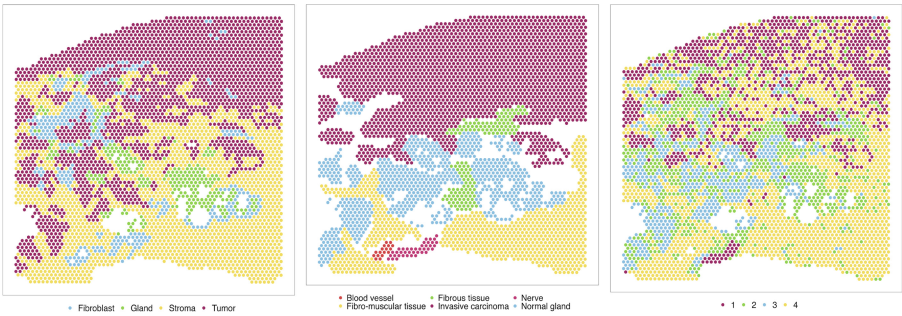


Fig. 1. Interferometric measures of land elevation speed on December 26, 2015 in the Phlegraean Fields volcanic area (Italy).

5 Satellite Monitoring of Ground Motion in the Phlegraean Fields

We will conclude our contribution presenting an application to the remote monitoring of ground motion in the Phlegraean Fields volcanic area in Italy. The latest advancements in differential interferometric processing have indeed resulted in techniques which continuously provide high-resolution ground displacement images with precision up to the centimeter or even millimeter scale [10]. In Fig. 1, an example of these reconstructed data at a given time stamp is shown. Such diffuse information on how the ground elevation evolves over time offers the opportunity for simultaneous monitoring of extensive areas susceptible to environmental risks. In detail, in this study, we will make use of Synthetic Aperture Radar data processed with multi-temporal differential interferometric techniques to build an automatic data-driven satellite-based real-time monitoring system of the area of the Phlegraean Fields, Italy, that can be used by public institutions and agencies to promptly and systemically monitor the seismic and bradisismic activities in the area.

Acknowledgments. This work is partially supported by ACCORDO Attuativo ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano and by MUR grant Dipartimento di Eccellenza 2023–2027. The authors also gratefully acknowledge the financial support of IREA-CNR (Istituto per il Rilevamento Elettromagnetico dell'Ambiente del Consiglio Nazionale delle Ricerche).

References

1. Ajroldi, N., Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal prediction bands for two-dimensional functional time series. Comput. Stat. Data Anal. **187**, 107821 (2023)
2. Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J.: A prediction interval for a function-valued forecast model: application to load forecasting. Int. J. Forecast. **32**(3), 939–947 (2016)
3. Cao, G., Yang, L., Todem, D.: Simultaneous inference for the mean function based on dense functional data. J. Nonparametric Stat. **24**(2), 359–377 (2012)
4. Chernozhukov, V., Wüthrich, K., Yinchu, Z.: Exact and robust conformal inference methods for predictive machine learning with dependent data. In: Conference on Learning Theory. Proceedings of Machine Learning Research, pp. 732–749 (2018)
5. Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Stat. Sin. **21**(4) (2011)
6. Diquigiovanni, J., Fontana, M., Vantini, S.: The importance of being a band: finite-sample exact distribution-free prediction sets for functional data. Stat. Sin. SS-2022-0087 (2023)
7. Diquigiovanni, J., Fontana, M., Vantini, S.: Conformal prediction bands for multivariate functional data. J. Multivar. Anal. **189**, 104879 (2021)
8. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer (2006)
9. Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of theory and new challenges. Bernoulli (2021)
10. Gabriel, A.K., Goldstein, R.M., Zebker, H.A.: Mapping small elevation changes over large areas: differential radar interferometry. J. Geophys. Res. Solid Earth **94**(B7), 9183–9191 (1989)
11. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 148–155. Morgan Kaufmann Publishers Inc. (1998)
12. Hyndman, R.J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: a functional data approach. Comput. Stat. Data Anal. **51**(10), 4942–4956 (2007)
13. Horváth, L., Kokoszka, P., Rice, G.: Testing stationarity of functional time series. J. Econom. **179**, 66–82 (2014)
14. Inselberg, A.: The plane with parallel coordinates. Vis. Comput. **1**(2), 69–91 (1985)
15. Lei, J., Rinald, A., Wasserman, L.: A conformal prediction approach to explore functional data. Ann. Math. Artif. Intell. **74**(1–2), 29–43 (2015)
16. López-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Am. Stat. Assoc. **104**(486), 718–734 (2009)
17. Ramsay, J.O.: When the data are functions. Psychometrika **47**(4), 379–396 (1982)
18. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics, 2nd edn. (2005)
19. Shafer, G., Vovk, V.: A tutorial on conformal prediction. J. Mach. Learn. Res. **9**, 371–421 (2008)
20. Sun, Y., Genton, M.G.: Functional boxplots. J. Comput. Graph. Stat. **20**(2), 316–334 (2011)
21. Torti, A., Pini, A., Vantini, S.: Modelling time-varying mobility flows using function-on-function regression: analysis of a bike sharing system in the city of Milan. J. R. Stat. Soc. Ser. C **70**, 226–247 (2021)

22. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005)

From Data-Driven to Expert-Guided: Combining Unsupervised and Semi-supervised Clustering in Spatial Transcriptomics

Andrea Sottosanti(✉), Sara A. Castiglioni, and Davide Risso

Department of Medicine, University of Padova, via Giustiniani 2, 35128 Padua, Italy
andrea.sottosanti@unipd.it

Abstract. One of the challenges in spatial transcriptomic experiments is identifying clusters of genes that exhibit similar expression patterns within specific regions of a tissue sample. The SpaRTaCo model, proposed by A. Sottosanti and D. Risso in 2023, offers a fully data-driven approach for the spatial classification of a tissue based on gene expression levels. Additionally, pathologist annotations of tissue samples are often available, albeit with significant variations between annotations and the data-driven analysis. In this work, we present a pivotal study focusing on a prostate cancer tissue sample. We demonstrate the integration of SpaRTaCo with two semi-supervised variants of the model, which incorporate external biological knowledge. This integration aims to uncover meaningful biological insights and specific gene expression patterns that may not be apparent through solely one of the two approaches.

Keywords: co-clustering · genomics · spatial correlation · spatial transcriptomics

1 Introduction

Spatial transcriptomics represents a groundbreaking category of sequencing technologies, offering the expression profiles of numerous genes in a tissue sample while preserving its spatial organization. By leveraging additional spatial data, researchers can gain deeper insights into the intricate biological mechanisms reliant on tissue cellular arrangements. Notably, the identification of spatially expressed (s.e.) genes—those displaying distinct spatial variation patterns—has emerged as a significant discovery [3].

Our recent work [2] introduced SpaRTaCo, a co-clustering model tailored for spatial transcriptomic analyses. SpaRTaCo effectively identifies s.e.g.enes active within specific regions of a sample, offering insights unattainable by existing methods. However, its computational demands are considerable, limiting its widespread application.

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 281–286, 2025.
https://doi.org/10.1007/978-3-031-64346-0_48

In addition to fully data-driven classifications, spatial experiments often entail manual annotations of the cellular composition by pathologists. These annotations serve as valuable sources of information that can enhance the inferential process. Although straightforward modifications enable SpaRTaCo to integrate such external data, it's essential to note that annotations may vary significantly depending on the expert, thus leading to potentially discordant results.

In this article, we present a pivotal study demonstrating the integration of results obtained from the original SpaRTaCo formulation with those derived from pathologist annotations. This combined approach can unveil specific discoveries and novel insights.

The rest of the manuscript is structured as follows. Section 2 details SpaRTaCo and two semi-supervised variants that incorporate external biological knowledge, while expediting the estimation process. Section 3 presents the analysis of a prostate cancer tissue sample using the three aforementioned methods, showcasing their efficacy in analyzing spatial transcriptomic data.

2 Methods

In this section, we introduce both unsupervised and semi-supervised clustering methods, which will be evaluated and compared in the next section using a spatial transcriptomic experiment.

Let $\mathbf{X}$ be the $n \times p$ matrix of a spatial experiment having the expression of n genes measured over p spots, whose spatial locations are known. We assume the existence of K gene clusters and R spot clusters, inducing a partition of the experiment matrix into $K \times R$ blocks. Thus, the kr-th block, has dimension $\dim(\mathbf{X}^{kr}) = n_k \times \ p_r$, and $\mathbf{X} = \left(\mathbf{X}^{kr}\right)$, with $k = 1, \ldots, K$ and $r = 1, \ldots, R$.

2.1 SpaRTaCo

The model introduced by [2] performs simultaneous clustering of genes and cells while considering the spatial correlation of the data. Given the clustering labels, let $\mathbf{\Sigma}_{kr} = \text{diag}(\sigma^2_{kr,1}, \ldots, \sigma^2_{kr,n_k})$ be a diagonal matrix containing the variances of the n_k genes assigned to cluster k. The model assumes that

$$\mathbf{X}^{kr} | \mathbf{\Sigma}_{kr} \sim \mathcal{N}_{n_k,p_r} \left(\mu_{kr} \mathbf{1}_{p_r}, \mathbf{\Sigma}_{kr}, \mathbf{\Delta}_{kr}\right), \qquad \sigma^2_{kr,i} \sim \mathcal{IG}(\alpha_{kr}, \beta_{kr}), \tag{1}$$

where $\mathcal{N}_{n,p}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ denotes the $n \times p$ matrix-variate normal distribution with mean matrix $\mathbf{A}$ and row and column covariance matrices $\mathbf{B}$ and $\mathbf{C}$, respectively. We express $\mathbf{\Delta}_{kr}$, the covariance matrix of the spots, as

$$\mathbf{\Delta}_{kr} = \tau_{kr} \mathbf{K}\left(\mathbf{S}^r; \phi_r\right) + \xi_{kr} \mathbb{1}_{p_r}. \tag{2}$$

In (2), the kernel matrix $\mathbf{K}(\cdot;\cdot)$ expresses the correlation of the gene expressions across the spots belonging to cluster r (with spatial coordinates $\mathbf{S}^r = \left(\mathbf{s}^r_j\right)$, and $j = 1, \ldots, p_r$). The (j, j')-th element is given by the exponential covariance function $\exp\{-||\mathbf{s}^r_j - \mathbf{s}^r_{j'}||/\phi_r\}$. The parameters τ_{kr} and ξ_{kr} quantify the amount

of spatial variation and residual intra-block variability, respectively. Moreover, we will make use of the quantity τ_{kr}/ξ_{kr} to measure the amount of spatial variability compared to the residual variability, and for this reason we call it *spatial signal-to-noise ratio*.

The model requires the estimate of the parameters and the clustering labels, both of the rows (genes) and of the columns (spots). We face the estimation problem by combining the stochastic EM and the classification EM.

2.2 Semi-supervised SpaRTaCo

Sometimes, external classification of the spots may be available. In such cases, we can integrate this external information to provide gene clusters based on biological annotations, while expediting the estimation process. Although the model structure remains consistent with Formulas (1)–(2), it no longer necessitates the allocation of spots into clusters. The estimation can now be executed using a simpler classification EM algorithm.

2.3 Sparse Semi-supervised SpaRTaCo

While the incorporation of external information streamlines the estimation process of SpaRTaCo, the semi-supervised variant discussed in Sect. 2.2 may still encounter challenges, particularly in high-dimensional experiments with a growing number of pixels. In addition to the previously mentioned versions of SpaRTaCo, we introduce a computationally efficient approach in this study. This approach leverages nearest neighbor Gaussian processes [1] to approximate $\mathbf{\Delta}_{kr}$ in (2) with its sparse version, denoted as $\tilde{\mathbf{\Delta}}_{kr}$. Notably, the complexity of $\tilde{\mathbf{\Delta}}_{kr}$ scales linearly - both in terms of storage and computational time - compared to the version utilized by SpaRTaCo. The sparsity of $\tilde{\mathbf{\Delta}}_{kr}$ is achieved by inferring spatial correlation using a subset of $m < p$ spots (we here considered $m = 20$). This enhancement addresses the computational burden associated with the high dimensionality of spatial transcriptomic experiments.

3 Key Study: Human Prostate Cancer Tissue Sample

We evaluate the performance of the three models described in Sect. 2 using a human prostate cancer tissue sample analyzed with the 10X-Visium technology. After excluding the unexpressed genes, the final, normalized dataset is made of 500 genes measured over 4366 spots.

In this analysis, we consider spots annotation provided by Dr. Esposito from the Veneto Oncology Institute, Italy, as well as the one made by experts from 10X-Visium[1] (shown on the left and centre of Fig. 1, respectively). In the rest of the manuscript, we will refer to these annotations with Annotation 1 and Annotation 2, respectively. While both classifications share certain characteristics, notable discrepancies are observed between them.

[1] https://www.10xgenomics.com/datasets/human-prostate-cancer-adenocarcinoma-with-invasive-carcinoma-ffpe-1-standard-1-3-0.

3.1 Retrieving the Spatial Tissue Morphology

We fitted SpaRTaCo with $K = 3$ gene clusters (selected using the ICL criterion, as described in [2]) and $R = 4$ gene clusters, aligned with the number of different cell types indicated in Annotation 1. We display the spot clusters in the right plot of Fig. 1. The algorithm ran for 1,500 iterations, but did not meet the convergence criterion outlined by [2]. This lack of convergence is evident, for example, in the spread spots assigned to cluster 4 in the tumoral region. However, SpaRTaCo has successfully identified key tissue features. Overall, our model accurately identifies cancer cells in the upper part of the image and stroma cells in the lower part. Unlike Annotation 1, we do not observe multiple isolated blocks of cancer cells in the left part of the image, which aligns with Annotation 2. Clusters 2 and 3 identified by SpaRTaCo correspond to regions composed of fibrous tissue and normal glands, respectively, as indicated in Annotation 2.

Moreover, SpaRTaCo detects a prominent feature at the bottom of the tissue assigned to cluster 1. While Annotation 1 does not mention this feature, Annotation 2 indicates the presence of a nerve in that location. This finding underscores the ability of our model to capture spatial tissue structure and guide experts in identifying specific tissue components.

3.2 Discovering Clusters of Spatially Variable Genes

We proceed to compare the clustering outcomes obtained with our model. For illustrative purposes, we demonstrate the validity of our proposal using the two semi-supervised versions of SpaRTaCo described in Sect. 2. Given that the results obtained with SpaRTaCo align more closely with Annotation 2, we apply our semi-supervised model to that annotation, setting $R = 6$.

Figure 2 illustrates the spatial signal-to-noise ratio estimated with the two models. Notably, both methods identify a cluster of genes that exhibit high spatial expression in areas corresponding to blood vessels and nerves. This observation is particularly evident in the results obtained by the Sparse semi-supervised

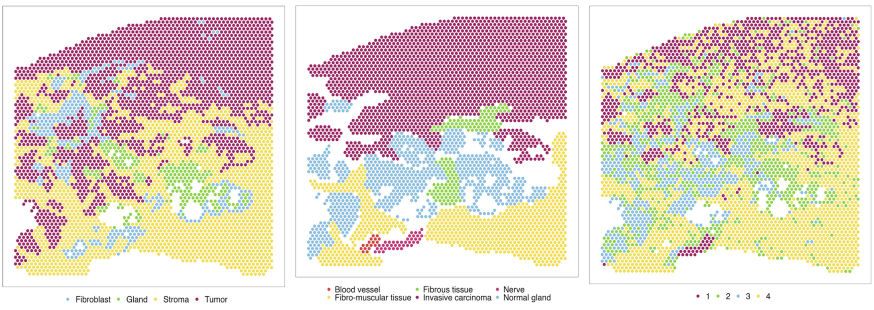


Fig. 1. Human prostate cancer tissue sample. Left: Annotation by Dr. Esposito. Centre: Annotation by 10X-Visium experts. Right: Clustering with SpaRTaCo. In the first two images, purple and yellow spots represent tumoral and stroma cells, respectively.

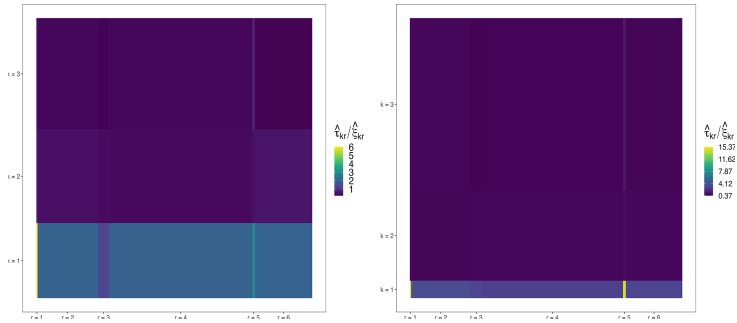


Fig. 2. Visualization of the experiment matrix $\mathbf{X}$ of the human prostate cancer tissue sample, partitioned into 3×6 co-clusters and coloured according to the spatial signal-to-noise ratios estimated with Semi-supervised SpaRTaCo (left) and Sparse semi-supervised SpaRTaCo (right).

version of SpaRTaCo (right plot), which assigns to cluster 1 genes with very high levels of spatial correlation in spatial regions $r = 1$ and $r = 5$. The remaining two clusters ($k = 2$ and $k = 3$) show poor spatial correlation in both models, indicating higher uncertainty associated with these clusters. Overall, the two model versions yield similar results and insights. However, the sparsity structure employed by the Sparse semi-supervised SpaRTaCo allows for better isolation of highly variable genes in specific tissue areas. Semi-supervised SpaRTaCo assign 134 genes to cluster 1, but only 30 of them are shared with cluster 1 of Sparse semi-supervised SpaRTaCo. These 30 genes show a very intense spatial expression level ($\hat{\tau}/\hat{\xi}$ up to 15.37) in the blood vessel and invasive carcinoma areas that appeared to be mitigated if considering also the additional 104 genes assigned to cluster 1 by Semi-supervised SpaRTaCo. Detailed results are shown with the confusion matrix displayed in Fig. 3.

From a computational standpoint, the sparse model significantly improved performance, requiring only 1.5 h for estimation compared to the 4.5 h needed for the Semi-supervised SpaRTaCo.

3.3 Detecting Single Highly Variable Genes

Once the model is estimated, we can sort the genes according to their specific variability, net of the spatial correlation of the data. This can be achieved through the conditional distribution of the random quantities $\sigma^2_{kr,i}$ appearing in Formula (1), given the data and the estimated model parameters. This step serves multiple purposes. First, it serves as an additional step for determining the presence of relevant expression patterns in space. Then, it identifies genes that exhibit specific patterns not captured by the clustering procedure. Lastly, it enables the identification and sorting of highly variable genes.

We provide two examples of how this phase can aid in discovering specific biological processes. First, we consider the SpaRTaCo model fitted as described

semi-supervised	Sparse semi-supervised 1	2	3
1	30	28	76
2	1	132	34
3	0	1	198

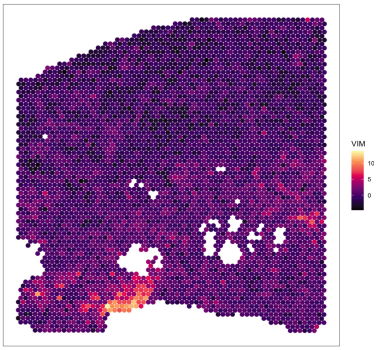


Fig. 3. Left: gene clustering obtained with the two semi-supervised versions of SpaRTaCo, estimated using Annotation 2. Right: spatial distribution of the gene VIM on the prostate cancer tissue sample. Colors denote the counts per pixel.

in Sect. 3.1. We previously noted that cluster $r = 1$ was compatible with the tumoral area identified by both pathologists (see Fig. 1). Analysis of highly variable genes revealed that, among the genes particularly active in this region, there are ACPP, AMACR, and APOD. These genes are associated with a high risk of prostate cancer and are often used as tumoral markers. ACPP is also active in clusters 2 and 3, but not in 4, which primarily consists of stroma cells.

The second example arises from the use of Semi-supervised SpaRTaCo fitted using Annotation 1. As previously mentioned, the pathologist who provided this annotation did not detect the presence of the nerve. Since we relied on the pre-given annotation, this additional component remained masked. However, through the analysis of highly variable genes, we discovered genes with very specific expression profiles in the area of the nerve. An example is provided by VIM, whose expression is displayed in Fig. 3. Thus, it is evident that this stage of the analysis serves to check whether the clustering procedure completely absorbs the spatial variability of the data or if there is something left that needs to be considered. The expression of VIM provides additional confirmation of the presence of the nerve in the bottom part of the image.

These results demonstrate the utility of using the SpaRTaCo model in all three versions mentioned in Sect. 2, and encourage further explorations and analyses using additional data and case studies.

References

1. Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E.: Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. **111**(514), 800–812 (2016)
2. Sottosanti, A., Risso, D.: Co-clustering of spatially resolved transcriptomic data. Ann. Appl. Stat. **17**(2), 1444–1468 (2023)
3. Svensson, V., Teichmann, S., Stegle, O.: SpatialDE: identification of spatially variable genes. Nat. Methods **15**, 343–346 (2018)

Author Index

A. Pollice and P. Mariani (Eds.): SIS 2024, ISSSAS, pp. 287–288, 2025.
https://doi.org/10.1007/978-3-031-64346-0